Internet Trolling: Social Exploration and the Epistemic Norms of Assertion

Daniel Munro

Penultimate draft; please cite version published in Philosophers' Imprint.

Abstract: Internet trolling involves making assertions with the aim of provoking emotionally heated responses, all while pretending to be a sincere interlocutor. In this paper, I give an account of some of the epistemic and psychological dimensions of trolling, with the goal of better understanding why certain kinds of trolling can be dangerous. I first analyze how trolls eschew the epistemic norms of assertion, thus covertly violating their conversation partners' normative expectations. Then, drawing on literature on the "explore/exploit trade-off," I argue that trolling is a kind of exploratory behaviour. Specifically, it involves exploring the consequences of violating an interlocutor's expectation that one will follow the epistemic norms of assertion. To defend this account, I argue that it explains various facts about trolls and the kind of pleasure they get from their activities. My account provides a deeper understanding of why trolling can be dangerous: namely, it explains why certain trolling behaviours contribute to radicalizing people into extremist or hateful ideologies, as well as why online platforms where trolls run amok gradually become polluted with extreme, hateful speech.

1. Introduction

Browse the comments section of any politically charged online article, social media post, or YouTube video, and you're likely to find attempts at "trolling." Trolls often post deliberately inflammatory content with the goal of provoking emotional responses. They aim to trick their targets into mistaking them for good faith interlocutors, thereby "baiting" them into responding in an emotional manner. This is typically for the troll's own entertainment, as well as the entertainment of anyone who happens to witness the exchange and recognize it as trolling.

Some instances of trolling seem mostly harmless, as when their contents aren't ethically problematic and no one takes the bait. However, trolling can also be dangerous. For one thing, empirical studies show that racist and misogynistic trolling can be part of a gradual radicalization into extremist or hateful ideologies (Munn 2019; Hoffman et al. 2020; Rauf 2021; Thorleifsson 2022). Furthermore, when problematic trolls are allowed to run amok, online platforms can gradually become cesspools of hateful speech. So, trolling can contribute to the degradation of both individual trolls' belief systems and broader online environments.

This paper gives an account of some of the epistemic and psychological dimensions of trolling, with the goal of better understanding the dangers of trolling that involves ethically problematic

content. I argue that trolling is often a form of *exploratory* behaviour, through which one explores the social consequences of replacing the usual epistemic norms of conversation with an aim of provoking emotional reactions. And I argue that this provides a deeper understanding of some of the potential dangers of trolling, both for individual trolls and for online communities.

§2 gives a basic account of trolling, drawing on existing literature and illustrative examples. §3 adds greater philosophical depth to this account by analyzing how trolling eschews the epistemic norms of conversation. §4 gives the paper's core arguments: after introducing the general idea of "social exploration," it argues that the way trolls eschew the epistemic norms of conversation enables trolling to function as a kind of social exploration. §5 argues that this account helps us better understand the dangers of trolling. §6 concludes by briefly sketching implications for online content moderation strategies.

2. Trolling: The basics

In popular usage, "internet trolling" is sometimes used very broadly, to refer to various antagonistic, deceptive, or humorous online behaviours (for attempts to taxonomize various uses of "troll," see Barney 2016; Cook et al. 2018; DiFranco 2020). In this paper, though, I'm focused on a narrower definition often used in academic work, one which better reflects how the term originated and how self-identifying trolls use it. On this definition, there are several necessary conditions an assertion must meet to qualify as trolling (my account in this subsection is especially inspired by Phillips 2015; Connolly 2022).[1]

Perhaps most central is that, when they post and comment online, trolls aim to provoke emotionally heated responses from their targets: anger, frustration, annoyance, shock, confusion, etc. Of course, one could accomplish this through many different kinds of speech. Crucially, though, trolls intend for their targets to interpret them as attempting to engage in sincere, good faith communication. An act of trolling is successful if the target "takes the bait" by interpreting the troll this way and responding in an emotional manner.

Consider an infamous example of successful trolling. On a 2008 episode of *The Oprah Winfrey Show*, Winfrey supported legislation aimed at combatting online child predators. Soon, trolls began

[1] As I do, accounts of trolling often focus primarily on examples involving *assertions*. One might object that other behaviours can count as trolling, such as certain ways of asking questions or posting images. If so, one can interpret my account as focusing on the specific subtype of trolling assertions. Regardless of exactly how we define trolling, I hope this paper demonstrates that this is an interesting kind of assertion worth considering in its own right.

posing as child abusers on the show's online message board. One posted: "WE DO NOT FORGIVE. WE DO NOT FORGET. OUR GROUP HAS OVER 9000 PENISES AND THEY ARE ALL RAPING CHILDREN!" This post implicitly references a couple of internet memes so is easily recognizable as a joke to fellow trolls. But Winfrey and her audience were successfully trolled when, on a subsequent episode, she followed up on her previous segment, saying:

> If you still don't understand what our children are up against, let me read you something which was posted on our message boards from someone who claims to be a member of a known pedophile network. It said this: "He doesn't forgive, he does not forget, his group has over 9,000 penises, and they're all raping children."

The audible gasps from Winfrey's audience indicate the troll's success: they interpreted him as posting in good faith, which elicited the intended shock.[2]

Online trolling humour often ranges from being in very poor taste (as in the previous example) to being intensely racist or misogynistic. To use another example: in 2010, when trolls began targeting posters on a Facebook memorial page for murdered teenager Chelsea King, one threatened to sexually assault the mother and sister of an online mourner; to the troll's delight, this threat was picked up and interpreted as a serious utterance by a local news station (Phillips 2015, 87). More recently, antisemitic trolling—for example, sharing conspiracy theories about politically powerful Jews aiming to exterminate the White race—has become increasingly common, enabled by alt-right platforms such as *The Daily Stormer* (Jakubowicz 2017; Kunzelman 2017).

Although the word "trolling" often brings to mind more extreme, ethically problematic examples, not all trolling is like this. For one thing, trolling can be more lighthearted. Suppose I have a film buff friend who likes to pontificate about her favourite auteur director. I might troll her by asserting a naïve critical assessment in a way that aims at provoking exasperation—something like, "He's okay I guess, but I found his last film pretty dull and pretentious." In a friendship context, such trolling can be a harmless bit of fun. Furthermore, even less lighthearted trolling can sometimes have positive consequences. One could go online and troll those who are themselves expressing racist or misogynistic views, which could derail a problematic conversation.

My descriptions of trolling in this section are meant to capture both problematic and more positive instances. However, my ultimate goal in the paper is to explain the potential dangers of trolling

---

[2] This infamous moment in 2000s internet culture is documented by Phillips (2015, 65-66) and the website *Know Your Meme*: https://knowyourmeme.com/memes/events/over-9000-penises.

that involves problematic content, such as the racist or misogynistic kinds. So, in what follows, I'll often use examples of this sort.

It might seem that trolling always involves asserting things one knows to be false or presenting oneself as believing things one doesn't believe. However, it's more accurate to say that trolling is simply *indifferent* to what's true or false, believed or disbelieved (cf. Phillips 2015, 26; Connolly 2022). Trolls don't intend to transmit either true or false information to their targets. Instead, they intend to provoke a certain kind of emotional response. One can often accomplish this by asserting something one knows to be false—for example, that one is part of a child abuse ring. One can also often accomplish this by asserting things one believes. Trolls targeting Chelsea King's Facebook memorial, for example, tried to provoke anger by posting irrelevant facts about her school grades and what she was wearing when she was murdered (Phillips 2015, 86). And a troll posting antisemitic conspiracy theories might believe these theories—it's just that, in posting them, his goal isn't to convince anyone that they're true.[3]

Trolling can seemingly be driven by a variety of motivations. Some are instrumental: some apparent trolls aim to take revenge on those who previously trolled them (Cook et al. 2018), while others are paid by private companies or governments to disrupt a competitor's online campaign or sow political confusion (Mahtani and Cabato 2019). However, there's controversy among self-identified trolls about whether those with purely instrumental motivations are *true* trolls. Some trolls take it to be constitutive of trolling that its primary aim is amusement at the target's heated reaction—amusement of the troll himself, and/or of anyone who witnesses the exchange and recognizes it as trolling (Phillips 2015, ch. 2).

Whether or not one thinks trolling is constitutively aimed at amusement, this at least characterizes *paradigmatic* examples of trolling. I'll restrict myself in what follows to such paradigmatic instances. That's because I'm interested in elucidating the nature of trolling *qua* behaviour that online users find pleasurable in and of itself. I take it that such pleasure is what originally caused trolling to become a popular online activity, one that emerged organically alongside the rise of the internet (cf. Phillips 2015, ch. 1). My goal is to better understand this behaviour to which many are drawn because

---

[3] In its indifference to the truth, trolling resembles Frankfurt's (2005) definition of "bullshitting" and Simpson and Michaelson's (2020) definition of "shilling." The key difference is that bullshitters and shills intend their interlocutors to believe their assertions, while trolls don't. We can also distinguish true trolling from what we might call a mere "trollish tone" (e.g., when politicians and political commentators speak in a snarky or condescending tone meant to rile up their audience, while still aiming to impart true beliefs).

they find it pleasurable, rather than because of external incentives. Such trolling can still have financial, ideological, or political goals as a secondary aim—there's likely a variety of psychological forces at work in any given instance of trolling (compare: one may become an artist primarily because of the joy they derive from it, but this doesn't rule out having a secondary financial motive). I mean merely to set aside instances of trolling guided only or primarily by external motivators.[4]

Strictly speaking, trolling needn't occur online. However, there are clear reasons why the internet is especially conducive to trolling. Successful trolling requires convincing targets that one is engaging in good faith. It's easier to deceive people about this online, since there are fewer opportunities to give away the fact that you're merely pretending to be a good faith interlocutor—you only have to pretend through text, not through speech, body language, etc. Furthermore, the relative anonymity of the internet allows one to avoid social sanctions that might result from being found out. While some of my arguments in what follows are applicable to offline trolling, they're intended to be read as focused on online trolling.

3. Trolling and the epistemic norms of conversation

As Grice (1989) observed, it's natural to think of a conversation as a cooperative activity in which participants expect one another to conform to certain norms. On this picture, these norms derive from the purpose or aims of a conversation. So, exactly which epistemic norms we take to govern assertions in general depends on how we construe the fundamental epistemic function of cooperative conversations.

I'll assume here that this function is to share *knowledge* amongst participants. So, following philosophers such as Williamson (2000), I'll assume that parties to a conversation are subject to the norm that they should assert P only if they know that P (the "knowledge norm of assertion," hereafter KNA). I'll formulate my arguments in terms of KNA because it's simplest to stipulate one such norm and stick with it; however, it should be possible to replace KNA in my arguments with some other norm (e.g., Grice's own view that one should assert only what one believes on the basis of adequate evidence; the view that one should assert only what is true; etc.).

Again, it's not merely that our assertions are governed by KNA unbeknownst to us. Instead, we're at least implicitly aware of this norm, which gives rise to certain expectations of conversational

[4] In calling those who troll for amusement the paradigmatic kind, I'm disagreeing with Barney (2016), who suggests internet trolls paradigmatically aim to disrupt ideological opponents' online communities. However, my focus is informed largely by Phillips' (2015) empirical investigations of troll culture.

participants: we expect one another to conform to this norm, and therefore to make knowledgeable contributions to the conversation (cf. Goldberg 2020; Westra and Nagel 2021). This at least characterizes how we expect things to go by default, when one has no prior, overriding expectation that one's conversation partner will be deceptive or misleading. That we have such a normative expectation explains various facts about how we evaluate other people's assertions. For example: it explains why it's legitimate to *challenge* an assertion by asking, "How do you know that?", a response which presupposes that what was asserted is known (Williamson 2000). Similarly, it explains why it's legitimate to *criticize* an assertion by exclaiming, "You don't know that!" (Kelp and Simion 2017), which suggests the assertion deviates from our usual normative expectations.[5]

When someone posts or comments online, she invites others to engage in conversation by responding. Trolls respond in a way that misleadingly presents themselves as good faith interlocutors. It therefore seems, intuitively, that we should be able to pinpoint some sense in which trolls violate epistemic norms, since they're doing something insincere in making their assertions. However, the epistemic problem with trolling doesn't straightforwardly reduce to violating KNA. Trolls often violate KNA by posting false content. But one can also successfully troll by posting contents one knows are true—by, for example, posting known facts about what a murder victim was wearing when she died, as in the Chelsea King example above. So, one can successfully troll even while complying with KNA. Even in such cases, though, it still seems intuitively that there's something epistemically problematic about trolling.

To see what, first note that we don't expect people to comply with KNA by merely happening to assert things they know, or by doing so accidentally. Instead, we expect people's assertions to be sensitive to the distinction between what they know and what they don't know, such that their assertions are *guided by* their knowledge. In other words, we expect people to assert things they know because they're actively *following* KNA and regulating their assertions accordingly. (For my purposes, we needn't specify exactly what this consists in, psychologically speaking—presumably, it typically involves an implicit awareness of KNA, rather than an explicit intention to follow KNA whenever one makes an assertion.) To make this point sharper, just consider someone who goes around making totally random assertions, but who ends up happening to assert things she knows on occasion. Such

[5] Of course, conversations can also have other functions (e.g., social bonding), and we apply other norms to assertions (e.g., politeness). The normative framework just sketched merely focuses in on the epistemic realm, bracketing others.

a person clearly violates the expectations we have of a cooperative conversation partner, even on occasions when her assertions accidentally line up with what she knows.

Our expectation that others assert things they know *because* they're following KNA, rather than accidentally, is normatively important. Someone who consistently violates it by intentionally regulating their assertions by a distinct policy is untrustworthy. Suppose I really want you to like me and be my friend, so I ignore KNA and intentionally conform only to a policy of asserting propositions that will flatter you and make you happy. The result may be that I mostly assert things I know ("I really like your new glasses!" "I bought you lunch on the way over!"). But any coincidence between my assertions and my knowledge would be accidental, rather than the contents of my assertions being sensitive to what I know. If you discovered my ulterior motive, you'd no longer trust me, because my assertions *could* come apart from my knowledge at any time, whenever doing so is a better way to garner your favour. Such a serial flatterer fails to be a cooperative conversation partner, even when she happens to comply with KNA by asserting things she knows.

So, in addition to merely expecting them to assert things they know, we also expect our conversation partners to assert things they know *because* they're following KNA and regulating their assertions accordingly. This is the expectation trolls violate, which is why their assertions are epistemically defective even when they happen to assert things they know. Like the serial flatterer described above, trolls are indifferent to KNA and intentionally comply with a distinct policy: to assert that which will provoke emotionally heated reactions. Of course, trolls also try to fool their interlocutors into thinking they're engaging in good faith. We can cash out "engaging in good faith" as, at least in part, aiming to comply with KNA. So, while trolls pretend they're attempting to comply with KNA, they intentionally comply with a distinct policy.

Now, one might wonder why I've chosen to characterize trolling so centrally in terms of its relationship to *epistemic* norms. Aren't there non-epistemic conversational norms that trolls more saliently violate, such as norms of civility or politeness? In fact, there's no systematic relationship between such non-epistemic norms and trolling. Trolls who are intentionally racist or misogynistic clearly violate such norms. However, as I argued above, trolling can be a more fun or lighthearted exchange between friends, and it seems wrong to classify such exchanges as necessarily uncivil or impolite. As this section shows, there's a much deeper relationship between trolling and violations of an interlocutor's epistemic normative expectations: while we expect each other to regulate our assertions by following KNA, trolling *by definition* involves regulating one's assertions by a distinct aim, the aim of provoking emotionally heated reactions.

This brings out another point of clarification: I don't mean to claim that eschewing epistemic norms and corresponding normative expectations is, in and of itself, *ethically* problematic. In more lighthearted trolling between friends, trolls still replace KNA with an aim of provoking an emotional reaction. As such, they still do something that violates our epistemic normative expectations of one another. This isn't necessarily ethically problematic, though, even if it is epistemically problematic—we sometimes allow epistemic goods to be outweighed by other goods, such as shared humour between friends. What makes the cases of trolling on which I'm primarily focused in this paper ethically problematic is their *content*—for example, their racist or misogynistic content—not merely their epistemic deficiencies.

Still, there's a close connection between the epistemic profile of trolling and the tactics some trolls use to avoid taking responsibility for posting ethically problematic content online. Specifically, the fact they're merely pretending to care about KNA gives trolls a kind of deniability when their racist or misogynistic posts are criticized for being false or not well-evidenced (e.g., because they originate from biases or problematic stereotypes). Such criticisms assume the poster cares about whether their assertion conformed to KNA, since they involve pointing out ways KNA was violated. Trolls can simply claim they were "just trolling," implying they weren't aiming to comply with KNA anyway, thereby dodging the criticism.

This section argued that trolls intentionally replace KNA with a distinct policy. The next section builds on this to argue that, by eschewing KNA in this way, trolling can become a way of engaging in social exploration.

4. Trolling as social exploration

I'll set aside trolling in §4.1, where I draw on literature on the "explore/exploit trade-off" to develop the general notion of social exploration. §4.2 then argues that trolling is one way of engaging in social exploration. §4.3 considers and rejects an alternative account, while showing how its insights are useful for further fleshing out my social exploration account.

4.1. Exploration, exploitation, and the violation of normative expectations

Any agent needing to acquire resources faces a trade-off between two modes of action: *exploiting* known sources of reward versus *exploring* novel sources of (potential) reward. Exploitation involves relying on one's existing mental model of the world to do things like access food and other resources; exploration involves gathering new information in order to refine or add to one's model. Exploration is costly in that it takes effort, and it's risky in that it brings one away from known sources

of reward. However, exploration has potential payoffs: it can allow one to learn about new and better sources. If a wild animal sticks closely to a known source of food, it may never learn that there's a richer source nearby. And if you always stick to the same neighbourhood restaurant, you may never learn that there's one you like better a few streets over.[6]

When engaging in exploration, it's not necessarily that one begins with *no* beliefs about the part of the world one is about to explore. Instead, exploration can involve acting on weak or vague expectations, then updating them. You might have an inkling that you'd enjoy Korean food, but not know which particular dishes you'd like best. So, you might act on this weak, vague expectation by trying out various dishes. In doing so, you might end up eating some you don't like. In the end, though, you can update your world model with a set of clear, precise beliefs about which Korean dishes you like best. Then, in the future, you can exploit this more refined model to always order your favourites.

It's thus optimal in the long run to sometimes trade exploitation for exploration. We should therefore expect evolutionary pressures to equip organisms with cognitive mechanisms that make them inclined to explore when they're not facing a scarcity of resources (Friston et al. 2015; Schulz and Gershman 2019). To pinpoint such a mechanism in humans, psychologists often appeal to the pleasure we derive from novelty: from learning new things, travelling, trying new foods, etc. (Dubourg and Baumard 2022, sec. 4). Since we take pleasure in such activities, we have a natural drive to seek novelty, thereby acquiring the epistemic gains afforded by exploration. So, to describe some action as exploratory is not to say that an agent necessarily has exploration in mind as an explicit goal that guides her actions. Instead, it's to give an evolutionary explanation for why the agent is equipped with psychological mechanisms that lead her to engage in that action, i.e., to describe the biological function of those mechanisms. The goal directly guiding the agent's action may be achieving the pleasure that comes from novelty. But the evolutionary reason the agent finds novelty pleasurable is because seeking out novelty is a way of engaging in exploration.

Now, examples of exploration often involve learning about one's physical environment, sources of food, and the like. However, we should also expect trade-offs between exploration and exploitation to arise in the social realm.

---

[6] The distinction between exploration and exploitation idealizes away from the fact that, in practice, the majority of actions involve both (e.g., exploiting your favourite, familiar restaurant involves learning new things about how busy it is that day, which staff are working, etc.; exploring a new environment still involves exploiting some prior knowledge about how the world works). Still, we can meaningfully speak about the *degree* to which an action is exploratory versus exploitative. When writing about exploration, I have in mind *primarily* exploratory actions.

Human survival depends in myriad ways on the complex social structures in which we're embedded. So, it can be beneficial to learn from exploring different strategies for interacting with others. If I have to trade with other humans for resources, I can stick to exploiting the same strategy for negotiating trades that I've always used; or, I can explore new strategies to see whether another is more effective. After a few years on the job market, I can continue to exploit the same job interview strategies I've used in the past; or, I can explore new strategies to see if they're more effective. And if you've always used the same dating app to find potential romantic partners, you can continue to exploit this strategy; or, you might explore alternative strategies (e.g., attending a speed dating event). In these cases, one has existing mental models of how certain types of interactions tend to unfold, and of how certain types of people respond to social strategies used in the past. However, one might have a hunch that some other strategy could work better. By testing it out, one can further refine one's overall world model.

Because social exploration can be beneficial, we should expect humans to derive pleasure from it just as they derive pleasure from exploring novel environments. And it does seem that it can be pleasurable to explore novel strategies for interacting with others: if you're stuck in a rut when it comes to job seeking or dating, there can be a thrill in going out on a limb and putting yourself out there in a new way.

In the rest of this subsection, I focus in on a specific kind of social exploration: exploration by intentionally eschewing conversational norms that an interlocutor expects one to follow.

As Theriault et al. (2021) argue, following norms that others expect you to follow is a way of exploiting a highly predictable social environment. That's because, when we comply with another person's expectations, their reactions in turn become more predictable to us. It's especially easy to see this with something like social norms of politeness. If I respond to another person politely, as they expect me to do, then we'll likely continue to have a relatively predictable exchange. If I instead respond with something unexpectedly rude or antagonistic, it becomes much less predictable how my conversation partner will respond, as well as how the rest of the interaction will unfold.[7]

[7] The claim here is just that, *on average*, following norms makes conversations more predictable, while violating norms makes things less predictable. This is consistent with thinking there are some cases where violating norms results in a highly predictable response (e.g., if I know someone always reacts in the same, predictable way to insults).

Something similar goes for the epistemic norms which we expect one another to follow, such as KNA.[8] There are many reasons that, on average, asserting only things that you know will make a conversation run more predictably. For mundane topics—e.g., introducing yourself, describing your hometown, discussing your academic interests—there's likely to be overlap with prior conversations you've had with past interlocutors, making it easier to predict interlocutors' responses based on past experience. Asserting things you know also makes it easier to predict and respond to questions about *how* you know those things, your justification for believing them, and the like, since we often retain facts about the sources of our knowledge (cf. Nagel 2015). And interlocutors' background knowledge about one another typically constrains which domains each expects the other to know about, thereby constraining expectations about which domains each is in a position to assert things about.

The fact that following conversational norms makes conversations more predictable means that one can engage in social exploration by intentionally eschewing these norms. For my purposes, one particular way of doing so will be most relevant: engaging in social exploration by intentionally replacing some norm with a distinct policy.

Consider again contexts like job interviews or dating. If your strategy always involves following KNA by being honest and sincere, your conversations might become relatively predictable after multiple experiences. One day, though, you might decide to try out a new strategy of saying only what you think will impress your interlocutor the most, even when it involves inventing facts about yourself. This will make it more difficult to predict your interlocutors' reactions. As such, it's a way of exploring a new strategy for interacting with others, one that (while ethically dubious) could potentially yield professional or romantic rewards. Something similar could be said of the serial flatterer discussed above. If you've always followed KNA in the past, you might one day shift to a policy of asserting only things that will make your interlocutors happy, because you want to see if this will win you more friends. This could be a way of exploring a new social strategy—a kind of "social experiment"—in virtue of the fact that you don't know exactly how others will respond.

As these examples show, it's possible to reap the rewards of social exploration by *pretending* to follow some norm, while instead replacing it with a distinct policy. In the job interview, dating, and

---

[8] It may be that norms of politeness are of a different kind from conversational epistemic norms, since the former are purely conventional while the latter are (as per Grice 1989 and various others since) rationally grounded. However, what they have in common is the psychological fact that having these norms in place, whatever their origin, generates *expectations* that our conversation partners will follow them. My analysis is primarily focused on these expectations and ways of violating them, rather than on (violations of) the norms themselves.

flattery examples, one tries to present oneself as a good faith conversation partner. Covertly, though, one is eschewing the norm one's interlocutors expect one to follow. That this sort of behaviour constitutes a kind of social exploration explains why it can be pleasurable. Think of those who derive pleasure merely from deceiving others: "catfishers" who enjoy creating fake online dating profiles, con artists who get a thrill from manipulating others, and the like. In these sorts of cases, engaging in deceptive social exploration becomes like a big game of pretend—at least, for the deceiver who is "in on" the fact that it's a game.

4.2. Trolling and social exploration

As per §3, trolling is one particular way of eschewing a conversational norm while pretending to try to comply with it—specifically, trolls pretend to try to comply with KNA while instead complying with a policy of asserting that which will provoke emotionally heated reactions. In the previous subsection, I argued that intentionally replacing KNA with a distinct policy, while pretending not to do so, can in general be a means of engaging in social exploration. This allows us to see how trolling could be one particular way of engaging in social exploration. And this would explain why trolling is an activity so many online users find pleasurable, given that we have a general psychological tendency to find pleasure in exploratory activities.[9]

This gives us an a priori, conceptual account of the sense in which trolling could be a means of social exploration. In the rest of this subsection, I argue directly that, in fact, internet trolls often are engaging in social exploration. I won't try to argue that this characterizes *every* instance of trolling—trolling can occur in many different kinds of contexts and interactions, and I don't think we can hope to give a uniform explanation of every case. However, I'll aim to give an account that applies to the behaviours of many paradigmatic internet trolls: those who self-identify as trolls and for whom it's a regular pastime to try to get a rise out of people on the internet.

The rest of this subsection argues that, by adopting this social exploration account, we can explain various facts about how trolls operate and the kinds of pleasure they derive from their activities.

[9] As per §4.1's explanation of exploration, this isn't to say that trolls have exploration or its epistemic benefits in mind as goals that explicitly guide them. Instead, it's to give an evolutionary explanation of why so many people take pleasure in trolling, where this pleasure directly guides them.

4.2.1. The demographics of trolls' targets

If trolling is characteristically a means of social exploration, it should often involve enriching or refining one's mental model of the world, as exploration in general functions to do. We can bring out how trolling does this by considering how internet trolls typically target groups of people who are socially or politically unlike them in some way.

We have sharper, more detailed expectations about how members of our own communities behave in social interactions, since we're more aware of the background knowledge and assumptions we share with in-group members. This makes it easier to predict how our conversations with such interlocutors will unfold. Conversely, we're less able to predict the behaviours of out-group members, since we're not as intimately aware of their background knowledge and assumptions. This means we stand to gain more from exploratory actions that engage with individuals from other social groups, which allow us to refine our beliefs about how outgroup members respond to different types of interactions and assertions.

Accordingly, internet trolls often target those who are unlike themselves. Right-wing trolls, for example, often target their political enemies (using trolling as way to "own the libs"—cf. Robertson 2021). Similarly, some who troll Facebook memorial pages claim that it's inappropriate to show earnest emotions in online, public settings, and they aim to poke fun at the type of person who would be willing to display their emotions this way (Phillips 2015, ch. 2). In both these cases, trolls with certain shared values use trolling as a way to interact with members of other social groups.

If trolling is often a means of exploration, this explains why such trolls target people who are unlike themselves: they stand to learn more from exploratory interactions with out-group members than with in-group members.

Note: I'm not claiming that, prior to trolling an out-group member, trolls are *totally* in the dark about how the target will react. As per §4.1's account, exploratory behaviours can begin from a place of vague or weak expectations, then function to sharpen or revise them. So, trolls might start with a hunch about how members of some population would respond to a certain conversational move, after which trolling allows them to sharpen their mental model of such targets. The pretend pedophile trolling Oprah on her show's message board, for example, might have a weak hope or expectation for an outraged reaction, then wait in anticipation to see how things pan out.

4.2.2. The demographics of trolls

The idea that trolling is often a means of social exploration also helps to explain why people with certain psychological characteristics tend to engage in trolling.

There's evidence that trolling is especially attractive to those who feel socially isolated or disenfranchised. Hong and Cheng (2018) found that feelings of inferiority are among the most significant predictors of online trolling behaviours. The same goes for depression, which they hypothesize stems from the fact that "depressed people have less intimacy and less personal control over conversation in daily interaction, are socially isolated, and have deficits in terms of social skills" (403). Similarly, Bor and Petersen (2022) found that various forms of online political hostility, including trolling, are correlated with a desire for greater social status. This data fits well with the nature of the fringe groups in which trolling tends to be especially popular—in particular, various extremist and hate groups, such as incel and white supremacist communities (see §5 below for more on such groups). Members of these groups often identify as social rejects downtrodden by society, and they often desire to escape offline social isolation and find a sense of belonging and community (Hoffman et al. 2020; Thorleifsson 2022).

We should expect those who feels socially isolated to be especially attracted to social exploration. For one thing, social isolation signals that one's current strategies for interacting with others haven't been paying off, meaning it would be beneficial to explore and discover new strategies. Furthermore, someone without much of a social circle will have had fewer chances to engage in social exploration than someone always surrounded by other people. So, the idea that trolling is a means of social exploration explains why people of this sort end up trolling.

4.2.3. The nature of the pleasure trolling elicits

My account of trolling as social exploration also explains certain facts about the particular sort of pleasure trolling elicits.

First, it explains why those who identify as internet trolls describe the paradigmatic aim of trolling as simply achieving the pleasure evoked by a target's heated reaction. Recall, from §2, that self-identified trolls argue that this is the true aim of trolling, rather than trolling being instrumental for fulfilling some practical goal. This is exactly what the social exploration account predicts.

To see why, first consider the kind of pleasure we derive from other exploratory activities. Again, exploratory behaviours in humans are driven by the pleasure we derive from *novelty*: from learning new things, travelling to new places, trying new foods, etc. Psychologists have also argued

that humans' attraction to *pretend play*, especially among children, evolved because pretense allows for exploration: it lets us explore novel spaces of possible actions and causal regularities, thereby sharpening our counterfactual reasoning skills and causal models (Gopnik 2020). Now, notice that all of these paradigmatic instances of exploratory behaviour have something in common: we're characteristically motivated to pursue these activities purely *because* they're pleasurable, rather than because they're instrumental to fulfilling some further goal. Exploring a new city and engaging in childhood play are intrinsically pleasurable, which is typically what drives these actions.

This lines up well with how trolls describe their own aims when they insist that their goal is merely pleasure and amusement in and of itself. The fact that pretend play, specifically, taps into the pleasures of exploration is especially revealing here. As I've argued, trolling involves a kind of pretense: namely, pretending to try to conform to the epistemic norm one is violating. Moreover, Phillips (2015, ch. 2) argues that trolling is akin to pretend play in that trolls typically adopt an online persona that's distinct from their "real life" personality: trolls often describe themselves as behaving online in ways they wouldn't in real life, with a disconnect between their true selves and their online, "playful" personas. Plausibly, then, the pleasure of trolling is much like the pleasure of childhood pretend play, which is pursued for its own sake.

Furthermore, my account of trolling as social exploration explains why trolling specifically elicits a *humorous* kind of amusement, from both troll and audience. Specifically, the fact that trolling involves exploring by violating normative expectations fits well with prominent "incongruity" theories about the nature of humour.

Incongruity theories are popular amongst philosophers and psychologists. They hold that humour involves some kind of incongruity between a situation and how we expect the world will or should be, given the norms governing our expectations. Carroll (2014) surveys instances of this. For example: it's funny to imagine a customer asking "to have his pizza sliced into four pieces rather than eight because he's on a diet" (20), since the customer violates logical norms of reasoning. Similarly, it's funny when Charlie Chaplin uses a person as an armrest, or a tablecloth as a handkerchief, since he violates our expectations about these objects' proper uses. Even more relevant for my purposes, Carroll (2014) points out that humorous incongruity can arise when people violate expectations for how conversations should unfold: "For example, conversational protocols are violated when in answer to the question 'Do you know what time it is?' one replies by simply saying 'Yes'" (21).

Trolling involves eschewing conversational epistemic norms with which we expect others to comply. If the troll is successful, his targets don't pick up on this violation, while the troll himself and

his knowing audience members do. (They may not explicitly conceptualize it as a violation of norms, but they at least notice it in the implicit way we notice norm violations in any instance of humour.) So, my account of trolling as social exploration can explain why a troll and his audience find trolling so humorous.

4.2.4. Trolling as pleasure in unpredictability

Pleasurable exploration requires an environment or social interaction that's unpredictable: if the results of our actions in some domain become totally predictable, we no longer derive pleasure from exploring it. Firsthand accounts from trolls suggest that trolling often involves taking pleasure in unpredictable outcomes of their interactions.

Cook et al. (2018) interviewed self-identified trolls from online video gaming communities, who regularly troll their teammates and opponents during gameplay. When asked about their motivations, one of the most common responses was that they start trolling other players once they become so experienced with a game that it becomes boring—once they've "seen all there is to see" within the game (3331). As Dubourg and Baumard (2022) argue, it's natural to think that exploring fictional environments within video games taps into our more general propensity to explore novel environments. The motivations trolls describe therefore suggest that, once gameplay has become too predictable and no longer yields pleasurable exploration, shifting to trolling is a way of making gameplay fun and exploratory once again.

Trolls also actively resist their own trolling activities becoming predictable. Online communities of committed trolls have at various times pushed back against attempts to make trolling more mainstream, organized, and formulaic (cf. Phillips 2015, ch. 8). One famous incident from 2009 exemplifies this.

At the time, online trolls were engaged in a campaign targeting the Church of Scientology. As the campaign wore on, it slowly began resembling a more organized, mainstream protest—including, for example, peaceful public demonstrations—instead of aiming to rile up the Church simply for trolls' amusement. Some committed trolls wanted to reverse this trend and re-focus on trolling for entertainment. They therefore initiated "Operation Slickpubes": they covered a shirtless man in a mixture of petroleum jelly, pubic hair, and toenail clippings, then sent him into some New York Scientology offices to smear as many surfaces as he could touch. Michael Vitale, one of the operation's instigators, later described it as demonstrating the troll community's willingness to engage in "any sort of motherfuckery" (Dibbell 2009).

Now, this isn't so much an act of trolling as it is simply an attempt to sow chaos—it's not an *assertion* that meets the definition of trolling given in §2. However, the message it sends is clear: trolls aren't interested in formulaic and predictable protests. It's relatively easy for the Church to anticipate the behaviours of peaceful picketers; these diehard trolls, though, believed that trolling should be undertaken in the spirit of Operation Slickpubes, which is far less predictable.

Trolls' pleasure in unpredictability also helps to explain why online trolling culture favours anonymity (over and above avoiding social sanctions, as per §2). The online spaces most notorious for trolling (e.g., anonymous message boards like 4chan, 8chan, and Reddit) allow users to conceal their identities. This allows trolls to increase unpredictability by creating new accounts and fake identities whenever they please, thus concealing their past track records of online posting. It's harder to predict when someone will troll in the future, as well as what sorts of content they'll post, when you can't see their past track record.

As per §4.1, if one's behaviours become less predictable to one's interlocutors, then these interlocutors' responses also become less predictable. Hence, one can continue to engage in social exploration by continuously subverting the expectations of one's targets.

4.3. Exploration or mere sadistic pleasure?: Accounting for trolls' personality traits

So far, I've argued that the thesis that internet trolling is often a means of social exploration explains various facts about trolls, regarding:

- The typical demographics of their targets (out-group members)
- The demographics of trolls themselves (those who feel socially dejected)
- The kind of pleasure trolling elicits (pleasure as end in itself; humour in incongruity)
- The ways trolls resist predictability

The fact that this thesis explains these various observations provides confirmation for it.

However, this confirmation isn't decisive if there's an alternative thesis that better fits the evidence. This subsection responds to perhaps the most obvious alternative: that trolling is merely a source of sadistic pleasure in manipulating, dominating, or otherwise harming others.

This hypothesis gets initial purchase from evidence from personality psychology, regarding the personality traits that correlate with enjoyment of online trolling. Studies show that trolls are especially high in *sadism* (i.e., they derive pleasure from harming others). There's also evidence that they're above average in *Machiavellianism* (which includes a tendency to manipulate others) and

psychopathy (which includes a lack of empathy and tendency to exploit others) (Buckels et al. 2014; Craker and March 2016; Sest and March 2017; Gylfason et al. 2021). Given all this, why do we need to appeal to social exploration to understand trolling? Why not just say that people do it because they enjoy the sadistic pleasure of manipulating others and using them as a means for their own enjoyment? Call this the "mere sadistic pleasure" account.

As I'll explain below, I grant that online trolls are often motivated by sadistic pleasure, and that this is part of a complete account of many trolls' behaviours. However, I'll first argue that this alternative explanation isn't sufficient on its own to account for many facts about trolls which, as I argued in the previous subsection, my social exploration account explains.

There are some facts that seem equally well explained by both accounts. This includes the fact that trolls tend to target out-group members, antagonizing their political opponents and perceived culture war enemies. It's plausible that evolution would equip us with psychological mechanisms that drive us to want to dominate or acquire power over out-group members, since there are obvious survival benefits from ensuring one's in-group stays dominant. Perhaps the sadistic pleasure many take in antagonizing perceived opponents and enemies is one such mechanism. In a similar vein, the mere sadistic pleasure account might also explain why those who feel socially dejected tend to be attracted to trolling: someone not used to feeling socially powerful might be especially attracted to opportunities to dominate others.

It's less clear that the mere sadistic pleasure account explains why trolling elicits humour as opposed to some other kind of pleasure. As per the previous subsection, my social exploration account explains this by appealing to the incongruity between trolling and the normative expectations we have of conversation partners. However, it's not clear why sadistically antagonizing out-group members should elicit a *humorous* response rather than some other kind of pleasure. There are other, twisted kinds of pleasure associated with sadistic or manipulative activities—think of the sorts of non-humorous pleasure or feelings of satisfaction some people derive from cyberbullying, cruelty to animals, or otherwise dominating others.

Unlike the social exploration account, the mere sadistic pleasure account also fails to explain why internet trolls tend to get bored and resist unpredictability when their activities become too predictable. If the pleasure of trolling comes merely from antagonizing or manipulating others, then there's no reason to think that trolls will want their interlocutors' reactions to be unpredictable. On the contrary, we'd expect them to figure out which strategies are effective for predictably provoking distress in their opponents, then continue to exploit these in future interactions. Once one figures out

what kinds of assertions tend to elicit strong emotional reactions from liberal internet commenters, for example, one could then stick with strategies known to be effective.

The social exploration account also unifies a wider range of cases of trolling than the mere sadistic pleasure account. Given my overarching aims in this paper, I have primarily focused on examples of internet trolling that involve harmful (e.g., racist and sexist) content. However, we shouldn't forget that trolling can also occur in more friendly contexts (e.g., when one gets bored while playing an online video game), where there might be less sadism involved. My social exploration account unifies both kinds of cases, since both can involve exploration via eschewing the standard epistemic norms of conversation. It does so while also helping us understand why internet trolls *tend to* come from certain demographics and prefer targeting certain kinds of people.

So, the mere sadistic pleasure account is, in various ways, less explanatorily powerful than the social exploration account. However, the social exploration account doesn't on its own explain the findings from personality psychology that motivated the mere sadistic pleasure account (i.e., why internet trolls tend to be higher in traits like sadism). This might still seem like a drawback.

Fortunately, however, there's a way to preserve the insights from both the social exploration and mere sadistic pleasure accounts. Specifically, findings about internet trolls' personality traits can help flesh out the social exploration account by explaining why some individuals turn to sadistic trolling over other forms of social exploration.

As per §4.1's discussion of the explore/exploit trade-off, we should expect humans to be equipped with a drive to explore novelty. However, as that subsection also brought out, there are *many* ways, besides trolling, of engaging in social exploration: replacing epistemic norms with a policy of flattering others; violating politeness norms; testing new dating or job market strategies; etc. So, the fact that trolling is a kind of social exploration is insufficient for understanding why some people choose sadistic trolling over other exploratory behaviours. To explain this, we can instead appeal to individual differences, including personality traits, that make sadistic trolling a more attractive form of exploration for some people. A more sadistic person, for example, might be more likely to engage in harmful trolling than testing out new kinds of job interview strategies. Findings about the traits typically exhibited by internet trolls can thus help us understand why some people find harmful, antagonistic trolling to be a particularly attractive form of social exploration.[10]

---

[10] One more finding from personality psychology might seem problematic for the social exploration account: namely, that there's no significant correlation between online trolling and *openness*

5. The dangers of trolling

In this section, I argue that my account of trolling as social exploration can help us better understand some of the dangers of ethically problematic internet trolling, such as trolling that includes racist or misogynistic content. Perhaps the most obvious danger is that it can harm its targets: by deceiving them, making them the butt of a joke, directing hateful language at them, etc. However, such harms to trolls' targets won't be my focus.[11] Instead, I'll first argue that my account helps to explain the epistemic dangers for trolls themselves; I'll then argue that my account helps to explain how this sort of trolling can degrade online environments in which it occurs.

Trolls' behaviours tend to escalate and intensify over time: they constantly look for new, more creative ways to troll their targets, in order to "keep things fresh" (Phillips 2015, 32). Empirical evidence shows why this escalation can potentially pose an epistemic danger: as these behaviours escalate, they can contribute to a shift from mere playful joking around to genuine beliefs in the contents one is posting.

This occurs in various extremist groups whose internet presences are steeped in ironic trolling—for example, the misogynistic incel community and various white supremacist communities. Activities like trolling are used as low-barrier ways to introduce potential recruits to these groups' ideologies. Newcomers often start out posting relatively mild content. And they may not initially believe hateful contents they post—when pressed, they often claim they're "just trolling" rather than sharing their beliefs (cf. Phillips 2015, 97). However, the epistemic danger is that, as their behaviours gradually intensify, this can contribute to the adoption of hateful beliefs (Munn 2019; Hoffman et al. 2020; Rauf 2021; Thorleifsson 2022).

Various psychological mechanisms might explain such shifts from mere online play to genuine belief. For example, it may be that acting "as if" one believes some hateful content results in cognitive dissonance when one doesn't actually possess the relevant beliefs, where resolving this dissonance involves revising one's beliefs to fit one's actions (cf. Guadagno et al. 2010; van Eerten et al. 2017). Or, it may simply be that posting more and more hateful jokes about members of some population

---

to experience, which involves seeking out novel experiences and information (Buckels et al. 2014; Gylfason et al. 2021). Doesn't the social exploration account predict that trolls are especially interested in seeking novelty? Actually, it doesn't: it's sufficient for this account that trolls are as attracted to exploration as the average person, since (as per §4.1) humans in general are attracted to exploration to some extent. These studies found that trolls exhibit about as much openness as the average person (i.e., they found neither a negative nor positive correlation between trolling and openness).

[11] On such harms, see DiFranco (2020) and Connolly (2022).

causes one to gradually dehumanize and "other" them (Munn 2019; Rauf 2021). Plausibly, various such psychological forces are at work in any given case. For my purposes, though, the key point is just the finding that becoming increasingly absorbed in hateful trolling can causally contribute to a troll coming to believe what he posts.

This means that, to fully understand how trolling contributes to online radicalization, we must understand why trolling behaviours tend to become more extreme over time. My account of trolling as social exploration explains this.

In general, we're attracted to exploring *novelty*. So, exploring a given physical or social domain becomes less pleasurable the more easily we can predict the results of our actions in that domain. In exploratory online interactions, then, one will derive less pleasure when one can predict how one's interlocutors will respond. This is why, as per §4.2.4, trolling often involves resisting predictability. However, trolling itself can become predictable if one trolls in the same way over and over again: after posting the same sort of inflammatory content enough times, one will start to be able to predict, based on past experiences, how others will react. To keep trolling pleasurable, one can post increasingly more extreme content, so that one can never fully predict the reactions it will solicit. This explains why problematic trolling behaviours escalate and become more intense over time, which then influences trolls' beliefs.

We can build on these insights about individual trolls to better understand how trolling degrades online environments in which it flourishes. Certain online platforms, especially those with very minimal moderation, are known as hotbeds of extreme, hateful trolling. This is true of, for example, fringe message board sites like 4chan and (the now defunct) 8chan. While trolling also regularly occurs on more mainstream sites like Facebook and Twitter, those sites don't have a reputation for being totally overrun with trolls, and their average user doesn't constantly encounter intensely misogynistic and racist trolling.

One simple explanation for why less moderated platforms feature more extreme rhetoric is that users gravitate to these sites out of a pre-existing desire to post such content. This is undoubtedly true. However, this explanation alone fails to fully account for the ways certain online communities evolved to become more and more extreme over time.

Consider message boards 4chan and 8chan. Founded in 2003, 4chan is a lightly moderated platform that's now infamous for hateful and extremist rhetoric. It's also synonymous with trolling, to the point that one can never tell whether users actually believe what they're posting. Within a decade of 4chan's original founding, the site gradually got more and more out of hand—despite their strong

commitment to light moderation, moderators eventually had to crack down on posts inciting violence and harassment. This was why 8chan was founded in 2013, intended to be even less controlled than 4chan (Chiel 2016). However, by 2019, 8chan was shut down: its security provider and domain host stopped supporting it after it became clear the site was helping to fuel offline violence, such as racially motivated mass shootings (Glaser 2019).

For a more specific case, consider the evolution of the incel community. In the late 1990s, the community began on a dedicated website, which was essentially a support group for young people experiencing difficulties dating and establishing fulfilling romantic lives. Over the next decade and a half, though, incel activities were slowly permeated with more and more hateful, trollish behaviour, especially after members migrated to lightly moderated sites like 4chan. Today, the community is rife with violently misogynistic trolling, including content encouraging violence against women and praising the perpetrators of gender-based mass shootings (Hoffman et al. 2020).

It's likely various factors contribute to the ways such communities become more and more extreme over time, such that the online platforms on which they make their homes eventually become so toxic. However, my account of trolling as social exploration provides at least a partial explanation. Because of the relationship between exploration and novelty, trolling behaviours tend to escalate over time to resist predictability—at least, if there are no incentives against this escalation, such as real-life social sanctions or online moderation. It therefore seems inevitable that relatively unmoderated online platforms will gradually become overrun with extremist trolling.

6. Conclusion

We expect interlocutors to be sincere, cooperative conversation partners. So, when trolls post racist or misogynistic content, targets often react as if responding to genuine, sincere expressions of belief (with anger, outrage, etc.). However, this social condemnation may not have the intended effect of causing trolls to stop posting; instead, it may just encourage them to explore novel ways of provoking more extreme reactions.

This puts other internet users in a bind regarding how to respond to hateful speech online. If we maintain our expectation that others are sincere interlocutors, and therefore respond as if they're expressing genuine beliefs, we risk pushing trolls to escalate their behaviours in increasingly extreme ways. However, if we assume everyone posting hateful content *could* be a troll, so refrain from reacting at all, then genuine expressions of hateful beliefs won't be called out.

My account thus suggests the responsibility for combatting hateful trolling should fall on moderators of online platforms, not individual users. Moderators should find strategies to block trolls from provoking heated reactions, thus preventing their behaviours from escalating. Whether this involves promptly removing or hiding hateful posts, or merely blocking other users from responding to them (by, e.g., disabling comments), moderation strategies should avoid causing trolls to explore newer, increasingly extreme tactics.

Acknowledgements: For helpful comments and discussion, thank you to David Barnett, Jennifer Nagel, Gurpreet Rattan, Julia Jael Smith, Zachary Weinstein, Seyed Yarandi, audience members at the Canadian Philosophical Association, students in my Fall 2022 "Technology and the Mind" class at the University of Toronto, and two anonymous reviewers for *Philosophers' Imprint*. This research was supported by the Social Sciences and Humanities Research Council of Canada, York University's Vision: Science to Applications project, and the Canada First Research Excellence Fund.

Bibliography

Barney, Rachel. 2016. "[Aristotle], on Trolling." *Journal of the American Philosophical Association* 2(2): 193-195.

Bor, Alexander and Michael Bang Petersen. 2022. "The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis." *American Political Science Review* 116(1): 1-18.

Buckels, Erin E., Paul D. Trapnell, and Delroy L. Paulhus. 2014. "Trolls just Want to have Fun." *Personality and Individual Differences* 67: 97-102.

Carroll, Noël. 2014. *Humour: A very Short Introduction*. Oxford University Press.

Chiel, Ethan. 2016. "Meet the Man Keeping 8Chan, the World's most Vile Website, Alive." *Splinter*. https://www.splinter.com/meet-the-man-keeping-8chan-the-worlds-most-vile-websit-1793856249.

Connolly, Patrick Joseph. 2022. "Trolling as Speech Act." *Journal of Social Philosophy* 53(3): 404-420.

Cook, Christine, Juliette Schaafsma, and Marjolijn Antheunis. 2018. "Under the Bridge: An In-Depth Examination of Online Trolling in the Gaming Context." *New Media & Society* 20(9): 3323-3340.

Craker, Naomi and Evita March. 2016. "The Dark Side of Facebook®: The Dark Tetrad, Negative Social Potency, and Trolling Behaviours." *Personality and Individual Differences* 102: 79-84.

Dibbell, Julian. 2009. "The Assclown Offensive: How to Enrage the Church of Scientology." *Wired*. https://www.wired.com/2009/09/mf-chanology/.

DiFranco, Ralph. 2020. "I Wrote this Paper for the Lulz: The Ethics of Internet Trolling." *Ethical Theory and Moral Practice* 23(5): 931-945.

Dubourg, Edgar and Nicolas Baumard. 2022. "Why Imaginary Worlds? the Psychological Foundations and Cultural Evolution of Fictions with Imaginary Worlds." *Behavioral and Brain Sciences* 45(e276): 1-72.

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

Friston, Karl, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. "Active Inference and Epistemic Value." *Cognitive Neuroscience* 6(4): 187-214.

Glaser, April. 2019. "Where 8channers Went After 8chan." *Slate*. https://slate.com/technology/2019/11/8chan-8kun-white-supremacists-telegram-discord-facebook.html.

Goldberg, Sanford C. 2020. *Conversational Pressure*. Oxford University Press.

Gopnik, Alison. 2020. "Childhood as a Solution to Explore–Exploit Tensions." *Philosophical Transactions of the Royal Society B* 375(1803): 20190502.

Grice, Paul. 1989. *Studies in the Way of Words*. Harvard University Press.

Guadagno, Rosanna E., Adam Lankford, Nicole L. Muscanell, Bradley M. Okdie, and Debra M. McCallum. 2010. "Social Influence in the Online Recruitment of Terrorists and Terrorist Sympathizers: Implications for Social Psychology Research." *Revue Internationale De Psychologie Sociale* 23(1): 25-56.

Gylfason, Haukur Freyr, Anita Hrund Sveinsdottir, Vaka Vésteinsdóttir, and Rannveig Sigurvinsdottir. 2021. "Haters Gonna Hate, Trolls Gonna Troll: The Personality Profile of a Facebook Troll." *International Journal of Environmental Research and Public Health* 18(11): 5722.

Hoffman, Bruce, Jacob Ware, and Ezra Shapiro. 2020. "Assessing the Threat of Incel Violence." *Studies in Conflict & Terrorism* 43(7): 565-587.

Hong, Fu-Yuan and Kuang-Tsan Cheng. 2018. "Correlation between University Students' Online Trolling Behavior and Online Trolling Victimization Forms, Current Conditions, and Personality Traits." *Telematics and Informatics* 35(2): 397-405.

Jakubowicz, Andrew. 2017. "Alt_Right White Lite: Trolling, Hate Speech and Cyber Racism on Social Media." *Cosmopolitan Civil Societies: An Interdisciplinary Journal* 9(3): 41-60.

Kelp, Christoph and Mona Simion. 2017. "Criticism and Blame in Action and Assertion." *The Journal of Philosophy* 114(2): 76-93.

Kunzelman, Michael. 2017. "Notorious Troll Calls the Online Tactic a 'National Sport'." *AP News*. https://apnews.com/article/technology-media-race-and-ethnicity-racial-injustice-hacking-04c73bb948ce4182845a6d27e0a9c3e1.

Mahtani, Shibani and Regine Cabato. 2019. "Why Crafty Internet Trolls in the Philippines may be Coming to a Website Near You." *The Washington Post*. https://www.washingtonpost.com/world/asia_pacific/why-crafty-internet-trolls-in-the-philippines-may-be-coming-to-a-website-near-you/2019/07/25/c5d42ee2-5c53-11e9-98d4-844088d135f2_story.html.

Munn, Luke. 2019. "Alt-Right Pipeline: Individual Journeys to Extremism Online." *First Monday* 24(6).

Nagel, Jennifer. 2015. "The Social Value of Reasoning in Epistemic Justification." *Episteme* 12(2): 297-308.

Phillips, Whitney. 2015. *This is Why we Can't have Nice Things*. MIT Press.

Rauf, Ateeq Abdul. 2021. "New Moralities for New Media? Assessing the Role of Social Media in Acts of Terror and Providing Points of Deliberation for Business Ethics." *Journal of Business Ethics* 170(2): 229-251.

Robertson, Derek. 2021. "How 'Owning the Libs' Became the GOP's Core Belief." *Politico*. https://www.politico.com/news/magazine/2021/03/21/owning-the-libs-history-trump-politics-pop-culture-477203.

Schulz, Eric and Samuel J. Gershman. 2019. "The Algorithmic Architecture of Exploration in the Human Brain." *Current Opinion in Neurobiology* 55: 7-14.

Sest, Natalie and Evita March. 2017. "Constructing the Cyber-Troll: Psychopathy, Sadism, and Empathy." *Personality and Individual Differences* 119: 69-72.

Simpson, Robert Mark and Eliot Michaelson. 2020. "The Big Shill." *Ratio* 33(4): 269-280.

Theriault, Jordan, Liane Young, and Lisa Feldman Barrett. 2021. "The Sense of should: A Biologically-Based Model of Social Pressure." *Physics of Life Reviews* 36: 100-136.

Thorleifsson, Cathrine. 2022. "From Cyberfascism to Terrorism: On 4chan/Pol/ Culture and the Transnational Production of Memetic Violence." *Nations and Nationalism* 28(1): 286-301.

van Eerten, Jan-Jaap, Bertjan Doosje, Elly Konijn, B. A. de Graaf, and Mariëlle de Goede. 2017. *Developing a Social Media Response to Radicalization*. WODC.

Westra, Evan and Jennifer Nagel. 2021. "Mindreading in Conversation." *Cognition* 210: 104618.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.