Conversational Goals and Internet Trolls[1]

Gretchen Ellefson

I. Introduction

Trolls are a virtually inescapable feature of the internet these days. Social media, comments sections, and nearly every other corner of the internet today are littered with instances of trolling, in which an internet user interacts with others online in a way that is designed to push buttons, draw out emotional responses, or undermine community dynamics in an online space. Despite how ubiquitous internet use is for communication, entertainment, community development, education, and nearly every other aspect of our lives, there has been relatively little philosophical engagement with the phenomenon of trolling,[2] and still less philosophical engagement that aims to address the distinctive linguistic and communicative behaviors of trolls.[3]

There are likely many reasons why trolling is more evident online than in face-to-face conversations, but one reason may have to do with adjustments in social norms and expectations that make it easier to engage in conversation in an uncooperative way: habit and socialization, norms of politeness,[4] and potential for awkwardness all recommend cooperativity (of a sort) over uncooperativity in face-to-face conversation. With the greater anonymity afforded by online communication, these considerations are progressively weakened to the point that in at least some online conversations, participants cease to behave in cooperative ways. As I'll argue, trolling behaviors illustrate a paradigmatic case of uncooperative communication in online spaces.

This claim stands in contrast to a view that is common among philosophers of language, according to which conversation is understood as an essentially cooperative practice. At least since Paul Grice's (1975) articulation of the Cooperative Principle, language theorists have discussed the relationship between conversation and cooperation, and often it is treated as a basic assumption that conversation is a cooperative activity. I hope to illustrate in the following pages that some trolling

[1] I am indebted to the Claremont McKenna College Department of Philosophy, and to the editors and contributors to this volume for their invaluable feedback.

[2] Exceptions include Barney 2016, Cohen, 2017, DiFranco, 2020.

[3] With the notable exceptions of Connolly, 2021 and Morgan, 2022.

[4] See Hardaker 2010, 2013 for a discussion of the role of (im)politeness in trolling.

behavior illustrates straightforward uncooperativity, since trolling involves of a rejection of the kind of goal alignment that is typically seen as characteristic of cooperative activities.

Note that I will be wary of drawing many categorical claims about trolls or trolling behaviors; I do not take myself to be arguing that *all* trolling is entirely uncooperative, nor that there are no normative guidelines for trolling.[5] However, one of the characteristic features of trolling behavior involves exploiting the conversational and non-conversational goals of others in a way that is, as I'll show, paradigmatically uncooperative. All in all, I will show that goal structure of trolling is one that is incompatible with expectations of cooperation in its various conceptions.

To establish these claims, I'll begin by clarifying how I am using the term "troll," which has become less precise as it has become more widely used. From common ways of characterizing trolling, I will aim to find a principled way of classifying internet users as trolls. I will do so by identifying a particular goal structure that is characteristic of trolling in which trolls aim to exploit and undermine the goals of their interlocutors. I'll then draw on common conceptions of cooperation, particularly those that relate cooperation to communication, and I'll show that trolling does not count as cooperative in any of the relevant senses. For all of the notions of cooperativity, the failure of trolling to count as cooperative ultimately comes down to the structure of the troll's communicative and non-communicative goals, and in particular, the relationship between the troll's goals and their targets'.[6] All notions of cooperativity that are put to use in linguistic theorizing rest on some relationship between the goals of conversational participants, which generally amount either to the goal of getting others to recognize your communicative intention, or else to engaging in some kind of coordination enabling participants to realize non-communicative goals. The goal structure of trolls is inconsistent with both of these cooperative goal structures. This is because the

[5] At the very least, trolls do sometimes cooperate with one another in the form of semi-coordinated efforts to troll specific communities. See section 2.2 in which I discuss the coordinated efforts of the Usenet newsgroup alt.tasteless to infiltrate rec.pets.cats, as well as the coordinated targeting of the Church of Scientology by internet trolls.

[6] I will use "target" or "target audience" to indicate individuals or groups that are the focus of the trolling. This is a more apt term than "audience" when speaking about trolls, since in online conversations, there are often other audiences besides the target(s): onlookers of various types, including other trolls who will be in on the joke. While this is an important feature of trolling, which Connolly (2021) addresses in his discussion, it is not my focus here. Lewiński and Dutilh Novaes (this volume) offer a useful model for this phenomenon.

troll generally aims not necessarily to communicate something particular to their interlocutor, but instead, to use their utterances to undermine the goals of their targets.

II. Internet Trolls

2.1 Definitions and the Role of Disruption

The following are all attempts to define 'internet troll':

> "A class of geek whose raison d'être is to engage in acts of merciless mockery/flaming or morally dicey pranking." (Coleman, 2010)

> "A troller is a CMC [computer mediated communication] user who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement." (Hardaker, 2010, p. 237)

> "Trolling is the deliberate (perceived) use of impoliteness/aggression, deception, and/or manipulation in [computer mediated communication] to create a context conducive to triggering or antagonizing conflict, typically for amusement's sake." (Hardaker, 2013, p. 79)

> "Online trolling is the practice of behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose." (Buckels, Trapnell, & Paulhus, 2014)

And a definition from within the troll community:

> "[T]rolls are the ultimate anti-hero, trolls fuck shit up. Trolls exist to fuck with people, they fuck with people on every level, from their deepest held beliefs, to the trivial. They do this for many reasons, from boredom, to making people think, but most do it for the lulz" (Encyclopedia Dramatica via Coleman, 2010)

The unifying theme between these definitions, I claim, is *disruption*. It will be useful for our purpose to think of this disruption in terms of the discourse or conversation in question: using impoliteness or aggression to disrupt an ordinary conversation or another person's ability to participate normally in conversation; "triggering or antagonizing conflict" either with another individual or within a shared online space, thereby disrupting the target's discursive aims; engaging in "mockery/flaming" or "morally dicey pranking" to disrupt the discourse with another individual or individuals, etc. These disruptions are often achieved when a troll exhibits "pseudo-sincere intentions" to participate in the online space, thereby enabling them to enter or remain in the conversation in order to disrupt it. Many of the definitions above highlight that these disruptions and conflicts are ultimately apparently pointless, or else the only or primary purpose is the amusement (or "lulz") of the trolls themselves. If we are to carve out a characterization of trolling that captures the phenomenon, a focus on pointless disruption seems to be a good place to start. I will point more specifically to precisely how trolls demonstrate this disruptive quality in section 3. First, it will be useful to give a bit more background on trolls and some examples of trolling.

2.2 Background and Examples

Whitney Phillips's (2015) ethnography of trolls is limited to discussion of "subcultural trolls" that are characterized by a "set of unifying linguistic and behavioral practices"(17) which were developed in online community spaces such as 4chan. Though this is a narrower category than my characterization will capture, subcultural trolls are arguably the original trolls, and their behaviors set the stage for contemporary trolling behaviors. One of the primary characteristics of the subcultural troll is the emphasis on "lulz," a sort of corrupted schadenfreude, in which the troll "celebrates the anguish of the laughed-at victim" (27). The primary motivation[7] for any subcultural trolling is to derive lulz at the expense of others. The pursuit of lulz is associated with one of the "Rules of the Internet," endorsed by the trolling community: that "nothing should be taken seriously" (Phillips 2015, 26; *Rules of the Internet*). Seriousness, earnestness, sentimentality, vulnerability, and naivete are

[7] Angela Nagel (2017) has objected that the real-life, violent actions that are encouraged on some message boards to be evidence that for some, trolling is not done only for the sake of lulz, but is for the sake of violence, especially misogynistic and racist violence (26-27).

all seen as violations of this rule. In response, trolls treat such expressions as an opportunity to troll for lulz.

This might play out in two ways: a troll might either identify vulnerability or emotionality in their target, or else they may aim to draw out an emotional reaction in a target. This emotional reaction might be offense, angst, or anger. Both approaches can be seen in different instances of RIP trolling, in which a troll makes outrageous, cruel comments about a deceased stranger in an online setting. In one example, a troll posted the following shocking comment on a YouTube video about the Christchurch earthquake:

> "I and the rest of the world are pleased your piece of shit family … are dead squashed filthy shit rotting in the ground. Especially those two filthy babies that were squashed REST IN PISS YOU FUCKING RODENT PIECES OF SHIT" (YouTube comment on a video about the Christchurch earthquake, quoted in Connolly 2021, 3)

In the next section, I will speak to the over the top, outrageous offensiveness of comments like this, and how it contributes effectively to the overall approach that trolls take. As we can already see, however, this comment is clearly aimed to antagonize the target audience. The disruption that this comment aims to cause can be thought of in a couple of ways. First, it might be understood as a disruption of the environment of orderly, predictably sympathetic comments on this specific video. But it could also be understood as an attempt to disrupt the process of national grieving more broadly. As Connolly (2021) points out, trolling of this sort leads to a dilemma for those who are engaging in good faith in the online conversation: either they ignore the comment, in which case there seems to be a tacit acceptance of this kind of behavior as acceptable and thus, the disruption successfully reframes and re-forms the state of the public conversation about the tragedy. On the other hand, they can directly address the commenter and the comment itself, which also allows for successful disruption, since doing so reframes the conversation *around the troll* (and thus plays very nicely into the troll's goal of disrupting the conversation) rather than around the tragedy.

In other examples, a troll may target a narrower audience. These cases may also be less overtly offensive than the example of RIP trolling above. Nevertheless, we can see the same pattern in which a troll makes a comment in order to disrupt the online community's discourse, and reframe it in a new (often less productive) way. In some cases, for example, a troll will join an online message board, page, group, etc. intended for members to discuss some particular shared interest. The troll will join the page apparently in good faith (demonstrating "pseudo-sincere intentions"), and at some point begin participating in a way that undermines the intended purpose of the online space. This is a tactic that trolls have used both on individual level and collaboratively, where trolls coordinate with one another to compound the disruption.

In an early example of a coordinated troll offensive aimed at community disruption, users of the Usenet newsgroup alt.tasteless coordinated an infiltration of the newsgroup rec.pets.cats, a group dedicated to discussions and image sharing of cats. The instigator of the offensive initially engaged the group by asking for advice about his fictional cats, whose names were internet slang terms for genitals (Phillips 2016, 18-19). After this, the instigator alerted alt.tasteless to his plan, and many other trolls flocked to rec.pets.cats to participate in the trolling, offering their own offensive and shocking content, including "tasteless advice," often suggesting that the author of the original post harm the cats (Quittner, 1994, discussed in Phillips 2015).

In this case, the disruption takes place on at least two levels: first, for any request for information or offering of advice in bad faith, the trolls waste the time and attention of good-faith participants in the group who attempt to offer genuine advice, correct bad advice, or otherwise engage with the trolls. Second, and more significantly, the coordinated trolling disrupts the basic functioning of the group. Whereas it was once functional for Usenet users who simply wanted to talk about cats, ask genuine advice about their cats, share updates about their cats, etc., the group became increasingly unusable for those purposes, since they are instead inundated with crude and offensive content.

Claire Hardaker's (2010, 2013) research into the linguistic and behavioral practices of trolls follows examples of this kind of bad faith engagement in a community board. She used a corpus drawn from the Usenet newsgroup rec.equestrian to identify features that users took to be an indication that a given user was trolling the group. Hardaker (2013) identifies a set of "strategies" that trolls use that all, to various degrees, disrupt the discourse taking place within the newsgroup. In one example, users on the board identify another user as a troll after they "ask for help with an unscrupulous horse dealer who supplied her with a six-month-old, untrained horse for her small daughter to ride" (Hardaker, 2013, 70). While the advice seeker did not admit to being a troll, and thus we cannot know that they were a troll rather than simply an especially naïve horse lover who is unaware of the risks of allowing young children near untrained horses, this is consistent with the disruptive element of trolling that I have been discussing: by soliciting unneeded advice, the troll engages in the group *as though* they are a sincere participant, effectively insinuating themselves into the community and gaining the ability to claim the time and attention of other users. However, by asking advice in a way that suggests extreme naivete, the troll raises frustration and anger from interlocutors (in this case, one interlocutor freely admitted to being "contemptuous" and "rude" to the troll, and expressed significant frustration at the presumed troll's naivete), further disrupting the productive exchange of horse-related ideas.

In many ways, these kinds of message board and newsgroup disruptions bear a family resemblance with political trolling. Political trolling can take many forms, but it often takes place within Facebook groups or on other sites where people gather with others with similar interests, especially political or ideological interests. Political trolls use these settings to disrupt the existing political aims or goals of the group, or else attempt to influence and adjust the group's aims or practices. Like the subcultural troll, the political troll aims at disruption. Instead of disruption for the sake of lulz, however, the political troll aims to alter group dynamics and goals as a whole, which is best obtained by getting other participants to agree to some view or aim.

There is a tension between the subcultural troll's commitment to apparently pointless disruption from which they derive lulz, and political or cause-oriented trolling. Indeed, this tension has more than once presented divisions between trolls. For example, when trolls from 4chan and Anonymous waged a trolling offensive against the Church of Scientology in early 2008, this marked a turning point for Anonymous, which we now tend to think of as a hacker/hacktivist collective. At its inception, however, it was merely a group of trolls, for whom exposing the Church of Scientology was an endeavor bound to yield a generous harvest of lulz. As the project, known as Project Chanology, progressed, some members of Anonymous became more motivated by moral concerns than they were by lulz. For others, this earnestness illustrated a betrayal of the basic commitment to lulz. The result was a rift in Anonymous between the subcultural trolls motivated by lulz, and the cause-oriented trolls who were motivated by moral or political concerns (Olson, 2012, Chapter 6; Phillips 2015 147-151).

Project Chanology may have been something of a bellwether for what was to come in the world of trolling. While earlier trolls at least claimed to be motivated primarily by lulz, it has become increasingly common to talk about trolling in terms of coordinated troll offensives with particular political, social, or ideological aims. In the US today, talk of "Russian trolls" immediately brings to mind coordinated efforts to infiltrate social media spaces to influence US political processes. Many different claims have been put forward regarding the goals of Russian trolls, but one hypothesis that seems to have some legitimacy is explained by Steven Wilson, a theorist of politics and social media: trolling offensives "exploit existing political fault lines like race and regionalism to increase polarization and disaffection with the political system," in turn undermining democratic institutions by disintegrating trust in those institutions and in other citizens (Wilson, 2020).

Given the way that the common conception of trolling has changed in recent years, a question arises about how we should think about trolling for the purposes of analyzing the behavior of trolls. If trolling historically has been conceptualized as lulz-oriented subcultural trolling, and now some people use 'troll' as a label for anyone engaging in bad or unfriendly behavior online, we have two options: either we can allow that anyone who is sometimes called a troll is in fact a troll, in which case it would make most sense to consider 'troll' a polysemous term, or we can attempt to construct a characterization of trolls that will allow us to give a principled reason to call some instances of bad behavior online "trolling," but not others. I will take the second option. I will propose a characterization of trolling in terms of the goal structure of trolls that will offer an intuitive and principled way of sorting trolling from non-trolling.

III. The goal structure of trolling

I have claimed that trolling centrally involves discursive disruption: either the disruption of an individual's aims in a particular online communicative context, or the disruption of the aims of a community. In both cases, we see that aims, or *goals* are centrally involved in trolling: the troll exploits and undermines the goals of others in pursuit of unshared goals that are not related in any direct way to the content communicated in the interaction or the intended purpose of the group. In posting an offensive comment on rec.pets.cats, a troll aims to realize a goal which is unrelated either to cats or to the specific content of the post. While political trolls' goals often bear *some* relationship to the purpose of the group, this goal is to undermine the group's intended purpose, rather than to contribute to it.

With these considerations in mind, I put forward the following as a characterization of trolling:

> The troll engages in online communication with one or more non-communicative goals that are unshared with, and often concealed from, their targets, that take precedence over any communicative goals, and that rely for their success on the disruption of the existing goals of other participants or the stated purpose or normal functioning of the online discursive community.

In defending this characterization, I will not defend the stipulation that "the troll engages in online communication." I limit my discussion to online communication simply because there are features of the internet that make it especially conducive to trolling – anonymity, lack of proximity to targets, lack of social repercussions, etc. – and because trolling is more common/obvious on the internet. That said, I am open to a more permissive characterization that includes non-internet trolling as well.

According to the characterization, the troll engages in online communication *with one or more non-communicative goals that are unshared with, and often concealed from their targets, that take precedence over the communicative goals of the troll.* The first part of this – that trolls have non-communicative goals that are unshared with their targets – is neither surprising nor deep, since virtually all conversational participants have some goals that are not shared with their interlocutors. It is the way that the troll prioritizes certain goals, and how these goals interact with the targets' goals, that distinguishes trolling from non-trolling exchanges.

In particular, these unshared goals "take precedence over the communicative goals of the troll." In most cases of communication, there are communicative goals that are at least mutually recognizable to all parties to the conversation; even if we don't know all of another party's goals, we know what they mean by their words, and what they are aiming to get their audience to understand from their utterances. These kinds of communicative goals serve non-communicative goals. For example, if I want you to close the window for me I may say, "It's getting a little cold in here," with the transparent intentions that 1) you will recognize that I have said that it is getting cold, and 2) that I am, through my utterance, making a request of you to close the window. That which I aim to communicate to you (that I find it cold; that I am making a request of you) are elements of my communicative goals. But in addition to these communicative goals, I ultimately have the goal that you do in fact close the window. This cannot be a communicative goal, since I cannot communicate the fact of your closing the window; it is up to you, not me, what you do with my utterance.

Nevertheless, it is a non-communicative goal that I pursue by getting you to identify some of my communicative goals.[8]

In the case of internet trolls, the tight connection between communicative and non-communicative goals, wherein communicative goals in some way serve non-communicative goals by getting our audience to see, first, what we are trying to communicate, and second, what we are trying to accomplish through these communications, is broken. The troll's non-communicative goals take precedence over any communicative goal to the extent that they need not really *have* communicative goals. At the very least, these communicative goals are not important except insofar as they serve the non-communicative goal. As Phillips puts it, "Trolls don't mean, or don't *have* to mean, the abusive things they say" (Phillips 2015, 26). Although Phillips is speaking of insincerity – that trolls' abuse need not actually match what they believe – I think that it may well extend to linguistic meaning as well. Although typically, the words and sentences troll produces have meaning, the troll's aims do not require that their audience successfully identifies those meanings; in fact, it is not clear that the trolls themselves need to have any particular communicative intention in producing a given utterance. They say things that they judge likely to have the effects they want to have, but if their utterances are misunderstood, if the target takes the utterance to mean something different from what the troll intended it to mean, this only presents a problem for the troll if it interferes with the troll's realization of their non-communicative goals.

A Refinery29 article gives an example of a comment posted on their website which illustrates this point well:

> Planned Parenthood has nothing to do with health pure birth control Some of the female gender don't know that there is a full proof way place a asprin between your knees each night and remove in the morning. (Norkin, 2017)[9]

[8] This characterization can be fruitfully spelled out in terms of Austin's (1962) distinction between locutionary, illocutionary, and perlocutionary acts. This treatment is also reminiscent of Grice's (1957) discussion of meaning. I lack adequate space to give a full account of how my characterization of trolling could be framed in terms of these prominent linguistic theories.

[9] Perhaps ironically for my purposes, the article which discusses this comment argues that we ought to move away from using the term 'troll' to describe all bad behavior online (a point with which I agree), including this comment. Nevertheless, the article quotes linguist Robin Clark, who describes this as a "deliberate attempt to derail a thread," which aligns nicely with my characterization of trolling.

Although there is a family of ideas that that this comment seems to reference, it is not clear that the comment actually *means* anything at all. At the very least, it does not have any one particular meaning that it would be possible for the author to intend for their audience to recognize; their disruption of the thread is meant to happen in part through the fact that their comment means many things and nothing at the same time. Further, even if this troll did have some particular intended interpretation in mind, it does not seem right to think of this "communicative intention" in the same way that we think of the communicative intentions in non-trolling situations: after all, if the targets misunderstand the troll's utterance, but the disruptive effect is still achieved, then the troll's communicative act has still been successful for the troll's purposes. In fact, the comment seems designed in such a way that nearly any interpretation, and any reaction, will count as a success.

This leads to the other element of the characterization: *The realization of the troll's goal(s) relies for its success on the disruption of the existing communicative or non-communicative goals of other participants or the stated purpose or normal functioning of the online community.* As we saw above, the reason why the communicative goals of the troll are subordinate to the non-communicative goals, to the degree that the communicative goals are nearly beside the point, is because the goal of the troll is a particular effect or set of effects on the target individual or community. These effects involve, to one degree or another, disruption. The deriving of lulz at the expense of others requires disrupting other parties' pursuit of their goals; the infiltration and transformation of a political group requires disrupting the purpose or normal functioning of that group, either to abandon certain ideals, adopt new ideals, or to take (or abandon) action in a way they otherwise wouldn't.

I take it that this disruption typically functions through the exploitation of the goals of others. The troll aims to use the goals they identify within a group or for a particular person and exploit or undermine those goals for their own purposes. A group like rec.equestrian has a general aim of offering advice and functioning as a community. This enables bad actors to exploit these helpful, pro-social attitudes and aims in order to undermine this community dynamic. It is only because people generally take other participants to be engaged in the same project that the troll can effectively take advantage of the participants in the group.

Of course, attempting to disrupt the goals of others is not something that is unique to trolling. In order to spell out what is distinctive about the mode of disruption we find in trolling, it will be useful to highlight the distinction between conflictual conversation and trolling. In conflictual communicative interactions such as interpersonal arguments, parties also in many cases aim to undermine and disrupt their interlocutors' goals. For instance, in an argument between partners about whether to make a new significant purchase, each party may be aiming to thwart the other party's goal of making or not making the purchase. In these cases, the other party's goal is the source of the conflict; the parties want different things, those things are not compatible, and so they face conflict. The partners may engage in an argument in which each party aims to push the other party to abandon their goal and switch sides; the parties each aim to realize their goals by getting the other to adopt their goal. In this sense, each party is aiming for the disruption of the other party's goal.

By contrast, in trolling, although the troll's and targets' goals are not compatible, the targets' goals are not an impediment for the troll. Instead, the targets' goals are the source of a vulnerability and thus an opportunity for the troll to take advantage of that vulnerability. The target has a sincere goal: perhaps having a nice conversation, or getting advice, supporting one another, or building political power, etc. The troll takes advantage of that, undermining the target's ability to effectively pursue their goal. This happens when the troll contributes to the social space in a way that redirects attention away from these goals and toward something else. Once this is accomplished, the target's goal is undermined, which itself contributes to the realization of the troll's goal. Yet this is not because the goals of the individual or community conflict in any direct way with the troll's goals; it is because the troll's goal is, simply, the disruption of the target's goals for lulz, political disruption, etc.

In the case of RIP trolling specifically, the disruption happens through shock value and offense; if the goal of the online space is to mourn, the RIP troll disrupts that goal with their shocking content, and often by redirecting the conversational space towards addressing the troll's comment. Yet this would not be effective if there were not an established goal that could be disrupted for laughs. In the case of political trolling or other cause-oriented trolling, the structure is similar: the troll's goal is served directly by undermining the targets' goals. In redirecting the conversation away from the initial goals of the community, or else by pushing the conversation in an increasingly radical or polarized direction, the targets' goals are undermined or significantly altered, thereby serving the goals of the troll. Further, it is the antecedent goals of the targets that the troll must use and exploit to reach their goals.

So far, we have mostly seen ways in which trolls might exploit goals that are not straightforwardly communicative goals: political goals, goals of community building, advice seeking, entertainment, etc. In some cases, trolls might also directly exploit the communicative goals of their interlocutors, for instance by deliberately misunderstanding a target's utterance (Paakki, Vepsäläinen, & Salovaara, 2021), or by focusing on a feature of the target's utterance besides the (usually clear) intended meaning. In such cases, the effect is to undermine the efficacy of the target's communicative goals by manipulating the conversation in such a way that the target's utterance has a different effect on the conversation than the target intended (and that, under ordinary circumstances, the target should be able to expect).

Paakki et al (2021) highlight deliberate misunderstanding as a common strategy of trolls. This does not necessarily mean that the troll interprets targets' utterances in ways that bear no resemblance to any plausible interpretation of the utterance. Instead, the troll interprets an utterance in a way that is clearly not intended, even though it might be a theoretically possible interpretation of the utterance under different circumstances. Consider this example of an overly literal interpretation in which A is demonstrating trolling behavior:

> A: There are two kinds of people in this world. People who live with cats and people whose houses don't reek of cat piss.
>
> B: Such a stunning insight. Did you come up with that all by yourself?
>
> A: The first part of the aphorism is quite common. The second part is an observation that a lot of people whose houses don't reek of cat piss tend to experience. So the answer to your question is yes and no. (Paakki et al 2021, 441)

In A's second utterance, they answer B's question as though it is intended *as a sincere question*. In this context, however, it is clear that B's question is intended as a kind of eye roll or a rebuke for a bad joke. Nevertheless, Paakki et al note that subsequent contributors to the conversation continue to engage with A as though they are a sincere, good-faith contributor to the conversation, rather than a troll. This disrupts the conversation by reframing it around troll's pedantic answer rather than the larger topic of the conversation.

IV. Cooperation

Having characterized trolling in terms of a goal structure according to which the troll realizes their goals only through undermining the goals of their targets, it should be intuitively clear why trolling is a good example of uncooperative communicative interaction. It is natural to conceptualize cooperation as something that involves the sharing of goals, and coordinating with respect to those shared goals. Not only is this an intuitive way of understanding cooperation, is it also the standard way of understanding cooperation within action theoretic accounts of cooperation: two or more parties count as cooperating only if there is some goal that they share and they are performing actions in order to (jointly) realize that goal (Searle 1990, Tuomela 2000, 2005, Gilbert 1989, 2014, Ludwig 2007, 2020).

While it should be clear that this action theoretic approach is not straightforwardly compatible with trolling, it is common in linguistic theorizing to start with the assumption that cooperation is in some way central to the practice of communication. The remainder of the paper will be devoted to considering some of the ways that cooperation has been treated as central to the practice of conversation in order to show that there is no sense of 'cooperation' in which it makes sense to conceptualize trolling interactions as cooperative. This shows, in turn, that communicative interactions in general is not necessarily cooperative.

4.1 Shared goals and essential conversational purposes

As I noted above, it is typical in action theoretic accounts to conceptualize cooperation as centrally involving the sharing of goals and/or intentions. In applications of this kind of account to linguistic or communicative contexts, some theorists have suggested that conversation itself has some kind of built-in goal. Whatever this intrinsic feature, anyone who engages in conversation must hold the intrinsic goal in order to count as engaging in conversation in the first place. Thus, any two conversational participants must share the goal. For instance, Stalnaker (2002) claims that it is a purpose that is "essential to the practice" of conversation that "people say things to get other people to come to know things that they didn't know before" (Stalnaker 2002, 703). According to this account, the sharing of information is "essential to the practice" of conversation, and so we can expect that one goal of participants within any standard conversation will be the sharing of information.

Yet this kind of an account actually highlights the uncooperativity of trolls: as noted above, the troll is not necessarily aiming to share knowledge; if they are misunderstood, or if their utterances are so obviously false that no one believes them, this is fine, provided that they succeed in disrupting an individual's goals or the general goals of a conversational space. In fact, it might be an effective tactic of a troll to say things that they expect to have no impact on anyone else's beliefs, either because their utterances are transparently false, or because they are so obvious as to be uninformative. This could be an effective way to raise the hackles of a target and derail a conversation.

Tuomela (2000, 155) considers and rejects an idea similar to Stalnaker's: that conversation includes the essential purpose or goal of achieving mutual knowledge. Yet, if we think that this essential purpose must involve shared intentions, he claims that it is implausible that mutual knowledge could count as the goal of conversation. This is because audiences in particular need not act intentionally, much less with any goal shared with a speaker, in order to understand a speaker's utterances. They often simply "take up" the information offered to them. In many cases, utterance interpretation is an automatic response to hearing a given utterance, rather than action that requires any kind of intention or goal, much less a shared goal.

4.2 Cooperation as a precondition for communication

A similar, though theoretically distinct, view is that cooperation is a precondition for communication. Ludwig (2020), for instance, claims that individuals "share an intention with respect to the use of a communicative system that allows them to talk" (Ludwig 2020, 25). To share a communicative system involves the use of "expressions for which [participants] share meaning conventions," and their use "in accordance with those conventions" (Ludwig 2020, 25).[10] That is, using language in accordance with conventions is what it is to cooperate with respect to a communicative system, and this is a necessary precondition for the very possibility of conversational communication. One must at least cooperate with the language itself (and by extension one's linguistic interlocutors) in order to communicate at all.

However, I do not find this a compelling way to show that cooperation is a precondition for communication; it is not clear that there is any meaningfully shared intention with respect to the use of a communicative system. First, communicative systems are not necessarily things about which we form conscious intentions; our communicative systems are typically largely automatic. When a person engages in a conversation with another, it is very rare that they consciously determine what communicative system they should use in the conversation. They are typically able to just start

[10] See also Asher and Lascarides's (2013) and Asher and Quinley's (2011) discussion of "basic cooperativity."

conversing. Thus, it is not clear in what sense these default modes are used *intentionally* and it is strange to think about cooperation in the absence of intention.

Even if we allow that intentions need not be consciously formed with respect to our communicative systems, however, I think it is the exception, rather than the rule, that these are *shared* intentions with respect to our communicate systems or meaning conventions. After all, shared intentions are intentions that are explicitly social: *I have an intention to Φ with A and A has an intention to Φ with me* or else, *we have an intention to Φ together.*[11] It is not enough that A and I happen to individually have the same intention. In the case of communication, then, it would seem that in each individual conversation I must form a new intention with each person with whom I engage to use the appropriate communicative system in that conversation. But this is implausible, especially in internet discourse, in which we often don't know who is involved in the conversation.[12]

To illustrate why shared communicative systems, or shared communicative goals in general, are not necessary in trolling specifically, consider again the comment from Refinery29, in which the commenter posted a comment that is ungrammatical to the point that is not possible to parse. The commenter in this case is intentionally *refusing* to use standard linguistic conventions. Moreover, this failure to observe linguistic conventions effectively serves the troll's purposes: it is disruptive in that it requires any reader to work harder in order to try to identify the intended meaning, and there are multiple ways one might feel pressure to engage: to clarify what is meant, to critique language use, or to object to the ideas one might guess the comment is intended to communicate.

In other cases, trolls intentionally use language that they do not expect their interlocutors to have access to, and so, intentionally use communicative systems that will be partially inaccessible for their targets. Recall that in the case of rec.pets.cats, we saw that one troll posted a question about what to do with their (fictional) cats, who they named using niche slang terms for genitals (Phillips 2016, Quintner 1994). The troll presumably did not expect that all readers would understand these terms. In fact, if the targets didn't understand the terms and subsequently used them in their responses, this is all the better for the troll's realization of their goal of lulz.

[11] See, for instance, Bratman 1992, 1993, Tuomela 2000, 2005, Gilbert 1989, 2014, and Ludwig 2007, 2020, for further discussion of the nature of shared goals or intentions.
[12] Thanks to Jennifer Saul for this point.

4.3 The Cooperative Principle without shared goals

I noted in the introduction that Grice's Cooperative Principle was influential in establishing the idea that cooperation is in some way central to communication. Further, some of the language Grice uses initially appears to suggest that the cooperation of the CP is akin to cooperation in the sense in which action theoretic accounts mean it. In particular, the CP states that conversational participants should make their contributions "such as is required…by the *accepted purpose or direction* of the talk exchange…" (Grice, 1975, 45; italics added). However, there is good reason to think that the notion of 'cooperative' used in Grice's CP may not involve the sharing of goals at all, despite the name of the principle and the reference to accepted purposes.

Although the language of the CP lends itself to the idea that there are shared goals between conversational participants, nothing in the CP (or in "Logic and Conversation" as a whole) explicitly specifies what Grice takes cooperation to be, or how precisely conversation counts as a cooperative activity. Nor is Grice explicit about what the "accepted purpose or direction" of a conversation must be. Thus, it is open to theorists to interpret 'cooperation' in this context in ways that do not resemble the shared-goal picture. As Davies (2007) argues, we must reconstruct Grice's sense of 'cooperative' from the way it functions in the Gricean system, rather than assuming it aligns with the typical use of the term.

Grice's primary aim "Logic and Conversation" is to outline an account of conversational implicature, the phenomenon by which we can communicate something more than or different from the primary, literal meanings of our words and sentences. The assumption of cooperativity is supposed to enable this. Grice gives the example of two people discussing a mutual friend:

A: Smith doesn't seem to have a girlfriend these days.

B: He has been paying a lot of visits to New York lately.

In this example, we are clearly meant to see that B is communicating that Smith does, or may, have a girlfriend in New York. Yet B has not said this explicitly. In fact, taken on its own, B's utterance is utterly irrelevant to A's statement. It is only because we *assume* that B means for their contribution to be cooperative, given "accepted purpose or direction" of the conversation, that we can infer what B is communicating.

Here we begin to see the role that cooperation is supposed to play in a theory of implicature: a hearer is able to derive the intended meaning of a speaker's utterance because they assume that the speaker's utterance is "such as is required…by the accepted purpose or direction of the talk exchange…" (Grice 1975, 45). On this view, the "accepted purpose or direction" does not play the role of a shared goal, but instead, that of constraining appropriate utterances; saying something unrelated to the conversation is "uncooperative" in that there is no way for the audience to work out what the utterance is meant to communicate. The audience, recognizing the "accepted purpose or direction" thus also recognizes the constraints on what utterances are appropriate, and can narrow down what the speaker must have meant by a given utterance.

Understood in this way, it makes less sense to conceptualize conversations as necessarily involving cooperation, per se. Instead, it makes more sense to think of conversation as reliant on participants *assuming* cooperation (in the Gricean sense) of their interlocutors. It may be tempting at this point to concede my claim that trolling conversations are not themselves cooperative since *conversations* need not be cooperative in themselves, but claim that trolling interactions nevertheless involve cooperation in the Gricean sense that participants must assume that their interlocutors' engagement is "such as is required." After all, trolling is most effective when a target takes the troll to be engaging sincerely and earnestly. In other words, the troll is dependent on the cooperativity of their targets in that they exploit the fact that their targets are cooperative, and thus will assume that the troll is cooperative. Thus, the Gricean analysis is actually helpful for describing trolling interactions.

I have two responses to this approach. First, I do not think that it is true that trolling always requires that the parties to the trolling exchange assume the cooperativity of the other party. Trolls cannot assume that their targets will always treat their utterances as cooperative. There are several reasons for this: trolls are more aware than anyone that there are other trolls who they might unwittingly try to target, and who will of course not assume their targets to be cooperative; trolls also know that many people on the internet these days come into any interaction with some amount of skepticism, assuming they will run into trolls; and in many online spaces, one does not know who will engage them, or, for that matter, whether their trolling will be seen by anyone at all, and thus they cannot assume that anyone will engage them cooperatively. Further, it is not necessary for a target to engage cooperatively in order for the trolling to be effective. A number of the examples that Hardaker (2010, 2013) discusses in her work involve targets identifying a user as a troll, and nevertheless getting pulled into engaging with the troll in a way that disrupts the conversation. They do not assume that the troll is making their contributions "such as is required," but they are nevertheless trolled.

My second response to this final defense of the relationship between trolling communication and cooperation is to note that even if were to concede that Grice's picture is always explanatory for trolling interactions (which of course I reject), I nevertheless suggest that the way in which trolls take advantage of the cooperativity of targets gives us very good reason to reject the framing that conversation, or conversational participants, are or must *be cooperative.* When mutual assumptions of cooperativity give rise to more effective communication, it makes sense to characterize this as a cooperative interaction. By contrast, when a troll exploits the cooperativity of a target, without reciprocating any assumptions of cooperativity, it makes much less sense to think of this as cooperative. Instead, the troll's behavior closely resembles that of the free rider, who takes advantage of a largely cooperative system without contributing cooperatively themselves, and who should therefore be understood as an uncooperative participant in the system.

V. Conclusion

I have shown that, if my characterization of trolls and their behavior is right, then regardless of how we understand 'cooperation', trolling does not typically count as a cooperative activity. If it is true that the practice of trolling involves prioritizing unshared, non-communicative goals in a conversation in a way that involves the exploitation of the goals of others and the treating of the troll's own communicative goals as secondary, then this is inconsistent with cooperation in any relevant sense. This shows that at least some conversations are not cooperative, and thus, communication does not require or presuppose cooperation, since some amount of communication (however imperfect) does still occur within trolling conversations.

Before I close this paper, one final point bears addressing: trolling is widely agreed to be in some way deficient as a mode of communicative and conversational engagement. Some may wish to respond by claiming that instead of denying the relationship between conversation and cooperation, we should instead deny that trolling counts as a conversation in the first place. Thus one could concede that trolls are not engaged in cooperative conversation, without thereby threatening the assumption that conversation is a cooperative activity. To make the claim that certain non-conversations are not cooperative does not establish that the assumption of cooperativity in conversation is false.

I hold, however, that it would be ad hoc to deny that communicative exchanges lacking mutuality count as conversations. I follow Green (1990) in thinking that we ought to be permissive with our understanding of what counts as a conversation: any kind of talk exchange, or "the purposeful use of natural language" is a conversation (Green 1990, 412). Trolling is indeed a purposeful use of natural language, and it is a kind of talk exchange in which communication can and does occur. I can see no principled reason for categorically excluding trolling exchanges from our understanding of 'conversation'. This is especially clear once we consider that, while we do associate some degree of mutuality with conversation, conversations vary enormously in the degree to which they are mutual.

Further, as Green (1990) and Davies (2007) note, the purpose of the CP and Grice's theory more broadly is to shed light on how communication happens, wherever it might happen. We will need to be able to derive implicatures in order to interpret utterances even when eavesdropping, when listening to radio broadcasts, or when watching movies. This suggests that the ability to interact mutually is not necessary for using the tools and resources provided us by the Gricean system.

If this is right, then we must consider trolling to count as conversation, or, at the very least, to be a kind of communicative exchange that should be subject to the same basic principles as any other communicative exchange. As we have seen, however, trolling is not a cooperative communicative interaction. Thus, we must abandon the assumption that cooperation is a precondition for or necessary feature of communication.

References

Asher, N., & Lascarides, A. (2013). Strategic Conversation. *Semantics and Pragmatics, 6*, 1-62.

Asher, N., & Quinley, J. (2011). Begging Questions, Their Answers, and Basic Cooperativity. *Proceedings of the 8th International Conference of Logic and Engineering of Natural Language Semantics*, 3-12.

Austin, J. (1962). *How to Do Things with Words.* Oxford: Oxford University Press.

Barney, R. (2016). [Aristotle], On Trolling. *Journal of the American Philosophical Association*, 193-195.

Bratman, M. (1992). Shared Cooperative Activity. *The Philosophical Review, 101*(2), 327-341.

Bratman, M. (1993). Shared Intention. *Ethics, 104*(1), 97-113.

Buckels, E., Trapnell, P., & Paulhus, D. (2014). Trolls just want to have fun. *Personality and Individual Difference, 67*, 97-102.

Cohen, D. (2017). The Virtuous Troll: Argumentative Virtues in the Age of (Technologically Enhanced) Argumentative Pluralism. *Philosophy and Technology, 30*(2), 179-189.

Coleman, G. (2010). Phreakers, Hackers, and Trolls and the Politics of Transgression and Spectacle. In M. Mandiberg, *The Social Media Reader* (pp. 99-120). New York: NYU Press.

Connolly, P. (2021). Trolling as speech act (or, the art of trolling, with a description of all the utensils, instruments, tackling, and materials requisite thereto: With rules and directions how to use them). *Journal of Social Philosophy*, 1-17.

Davies, B. (2007). Grice's Cooperative Principle: Meaning and Rationality. *Journal of Pragmatics*, 2308-2331.

DiFranco, R. (2020). I Wrote This Paper For the Lulz: the Ethics of Internet Trolling. *Ethical Theory and Moral Practice, 23*(5), 931-945.

Donath, J. (1999). Identity and Deception in the Virtual World. In M. Smith, & P. Kollok, *Communities in Cyberspace* (pp. 29-60). New York: Routledge.

Gilbert, M. (1989). *On Social Facts.* Princeton: Princeton University Press.

Gilbert, M. (1990). Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies in Philosophy, 25*, 1-14.

Gilbert, M. (2014). *Joint Commitment: How We Make the Social World.* Oxford: Oxford University Press.

Green, G. (1990). The Universality of Gricean Interpretation. *Proceedings of the Sixth Annual Meeting of the Berkeley Linguistics Society*, 411-428.

Grice, P. (1957). Meaning. *The Philosophical Review*, 66(3), 377-388.

Grice, P. (1975). Logic and Conversation. In P. a. Cole, *Syntax and Semantics 3: Speech Acts* (pp. 41-58). New York: Academic Press.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, 6*(2), 215-242.

Hardaker, C. (2013). "Uh....not to be nitpicky,,,,,but...the past tense of drag is dragged, not drug": An overview of trolling strategies. *Journal of Language Aggression and Conflict, 1*(1), 58-86.

Ludwig, K. (2007). Collective intentional behavior from the standpoint of semantics. *Noûs, 41*(3), 355-393.

Ludwig, K. (2020). What is Minimally Cooperative Behavior. In A. Fiebich, *Minimal Cooperation and Shared Agency* (pp. 9-40). Cham: Springer.

Morgan, A. (2022). When Doublespeak Goes Viral: A Speech Act Analysis of Internet Trolling. *Erkenntnis*.

Nagel, A. (2017). *Kill All Normies: Online culture Wars from 4chan and Tumbler to Trump and the Alt-Right.* Winchester: Zero Books.

Norkin, L. (2017, March 28). No, That Shitty Comment Isn't 'Trolling'. *Refinery29*.

Olson, P. (2012). *We Are Anonymous: Inside the Hacker World of LulzSec, Anonymous, and the Global Cyber Insurgency.* New York: Little, Brown, and Company.

Paakki, H., Vepsäläinen, H., & Salovaara, A. (2021). Disruptive online communication: How asymmetric trolling-like response strategies steer conversation off the track. *Computer Supported Cooperative Work, 30*, 425-461.

Phillips, W. (2015). *This is Why We Can't Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture.* Cambridge: The MIT Press.

Quittner, J. (1994, May 1). The War Between alt.tasteless and rec.pets.cats. *Wired.*

Rules of the Internet. (n.d.). Retrieved April 2022, from Encyclopedia Dramatica: https://encyclopediadramatica.online/Rules_of_the_Internet

Sarangi, S., & Slembrouck, S. (1992). Non-cooperation in communication: A reassessment of Gricean pragmatics. *Journal of Pragmatics, 17*(2), 117-154.

Searle, J. (1990). Collective Intentions and Actions. In P. Cohen, J. Morgan, & M. Pollack, *Intentions in Communication* (pp. 401-415). Cambridge: The MIT Press.

Stalnaker, R. (2002). Common Ground. *Linguistics and Philosophy 25*, 701-721.

Tuomela, R. (2000). *Cooperation: A Philosophical Study.* Dordecht: Kluwer Academic Publishers.

Tuomela, R. (2005). Two Basic Kinds of Cooperation. In D. Vanderveken, *Logic, Thought, and Action* (pp. 79-107). Dordrecht: Springer.

Wilson, S. (2020, October 22). *What are Russia's goals with disinformation on social media*. Retrieved February 2022, from Brandeis Now: https://www.brandeis.edu/now/2020/october/elections-russia-disinformation-social-media.html