**Unintentional trolling: How subjects express their prejudices through made up stories**

**[please find the published open-access version at** [**https://link.springer.com/article/10.1007/s13347-020-00425-4**](https://link.springer.com/article/10.1007/s13347-020-00425-4) **]**

**1. Introduction**

In this paper, we will examine disruptive behavior on social media platforms, such as Facebook. This behavior is often referred to as “trolling,” which is typically considered as an intentional behavior. Drawing on data from the Wikipedia platform, Shachaf and Hara (2010) define trolling as: “repetitive, intentional and harmful actions” (p. 363). By intentionally breaking certain rules of conduct, trolling damages the platforms on which it takes place, and the people that repeatedly perform these disruptive acts are “trolls.” As motivations for trolling Shachaf and Hara list boredom, the search for attention, fun, and a desire for revenge. Such behavior causes harm to an online-community and might lead to its destruction. This can happen, for example, if an actor harms the online community whose political views they do not share.

Phillips (2016) detailed study on the trolling subculture describes trolling as an intentional activity, too. Phillips describes the values and mechanisms at play: the objective is that of “lulz,”, which is the enjoyment of other people’s rage or embarrassment. Hunting for *lulz* is the goal of trolling behavior and a justification for it (hence the wide-spread phrase “I did it for the lulz”). One important mechanism of trolling is the use of memes: they conserve and share the gained *lulz*.[[1]](#footnote-2) For instance, if a person reacts angrily because they are a victim of trolling (i.e., *lulz* was achieved by the troll) then the reaction can become memetic material; trolls might then use this to create even more *lulz*. Knowing these memetic references helps to constitute an exclusive circle of informed peers, who possess insider knowledge. Phillips (2016) clarifies that trolls know what they are doing, they are following long-term plans and are well organized. Accordingly, these trolls act *intentionally*.[[2]](#footnote-3)

Psychologists, however, commonly characterize trolling without referring to the trolls’ intentions. In contrast, they highlight distinctive behavioral patterns: “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 et al., 2014, p. 97). In this paper, we adopt this definition because it fits with the evolving nature of trolling: as Phillips (2016) points out, the trolling phenomenon has changed and has become more heterogeneous. Trolling is no longer an exclusive activity of an elite circle of trolls, because the concept of trolling became known to the public. This led to a blurred understanding, as different types of destructive online behaviors are now called “trolling” too.   
For this reason, we argue that there is a type of trolling that is unintentional, and that the actors in these cases have no malicious intentions. Also, we question the trolls’ alleged *irrationality* even though they appear to be resistant to facts. Therefore, we will analyze typical and characteristic behavioral patterns of these actors, such as creating and spreading *fake news* on social networks like Facebook or Twitter.[[3]](#footnote-4) For instance, in Germany there are a vast number of false reports that discredit refugees. The online project HOAXmap (https://hoaxmap.org) tries to keep track of this flood of false information, listing hundreds of fake news reports and providing information to disprove them. But even if the police or other agencies detect such false reports, the people spreading and believing fake news often show a *resistance to facts*: many people ignore the data that speak against their original assumption. The comment sections on Facebook are filled with agents that show such a resistance to facts. Here, one quickly gets the impression that arguments are only apparently exchanged.[[4]](#footnote-5)

To examine unintentional trolling behavior, we focus on the analysis of misunderstandings in communication. Based on empirical evidence, philosophy of psychology, and action theory, we argue that *unintentional* trolling exists. Thereby, we present a justified empirical hypothesis regarding this phenomenon. In addition, we outline criteria that could be used to identify unintentional trolls and to estimate their number.

If, as we shall argue, unintentional trolling exists, this may have consequences for moral and legal concerns. Therefore, we think that distinguishing between different types of trolling is important. Agents are considered morally responsible for their behavior when they act freely and knowingly. Here, “freely” refers to control conditions that are closely related to intentional behavior. Hence, when a person trolls unintentionally, they may not be blameworthy. Regarding legal issues, unintentional trolling could make a difference as well. In Arizona, the house bill 2549 (Arizona State Legislature, 2012) stated that the *intention* to annoy and offend is already sufficient for taking legal action. Although the bill was stopped, it shows that the actor’s intentions make a difference for legal matters.

In section 2, we focus on distinctive behavioral patterns, and characterize them as unintentional, but goal-directed. After introducing basic concepts for describing intentional action, we will illustrate our analysis with an example of unintentional trolling. We assume that the behavior in question is *unintentional* because no adequate personal-level goal-representation caused it. Language behavior is descriptive or expressive; the latter means that a subject articulates an affect, while the former means that a subject describes a fact. Stepping on a nail and shouting “Ouch!,” for instance, is an expressive use of language, which has no truth conditions, in contrast to descriptive language use. If a subject uses language descriptively, the subject describes facts, and the description can be true or false. As we shall argue, language can be expressive, even though it *appears* to be descriptive. This constitutes unintentional trolling: Unintentional trolls *seem* to use language descriptively while they, in fact, unintentionally express their prejudices.

In section 3, we present an argument for the assumption that utterances that appear to be descriptive can be expressive. We base our claims on the justification-suppression model (JSM; Crandall & Eshleman, 2003), which is an empirical framework for the analysis of self-reported explicit racial prejudices. Broadly speaking, the model holds that prejudices by themselves motivate subjects to express them; however, external (e.g., social norms) or internal (e.g., personal values) factors can suppress the expression of prejudices. As a consequence, subjects find a way to express their prejudices via justifications that bypass these suppressions. Against this background, we argue that unintentional trolling is a strategy used to express prejudices and bypass external social pressure.

In section 4, we argue that this type of trolling is heuristically generated and *goal-directed,* even though it is unintentional. We suggest that the behavior serves to overcome certain restrictions on participation in public discourse. We use the psychological concept of *storytelling* to illustrate how spreading false reports facilitates the expression of prejudice and influences the public discourse. Storytelling is a justification, like suggested by the JSM, that makes prejudice expression possible, while circumventing social sanctions.

In section 5, we outline behavioral characteristics of unintentional trolling that can distinguish unintentional from classic trolling. Although we cannot assess the mental states of subjects, we assume that the different psychological setup of intentional trolls will lead to different behavioral patterns that are distinct from those of unintentional trolls. In principle, these criteria can be used to estimate the number of unintentional trolls.

**2. Analysis of unintentional trolling**

The aim of this section is to develop an overall sketch of unintentional trolling. We first introduce our basic characterization of intentional behavior. Afterwards, we give an example of the behavior that we take to be unintentional trolling. After analyzing this example, we give a first characterization of unintentional trolling and discuss various subsequent questions regarding our definition in the following sections.

**2.1 Introducing relevant concepts**

First of all, we do *not* discuss whether the analyzed type of trolling relies on speech-acts (Austin, 1962). A speech-act plays a certain *social role* and can “create” facts on its own, so to speak. A baptism entails a speech-act because a new-born receives a name due to social conventions. Because of these social conventions, a speech-act can fail if the actor breaks certain social rules. For example, not anyone can perform a baptism. In this text, we focus on *acts of speech*: an act of speech is any act of uttering meaningful words (Green, 2017).As we are claiming that unintentional trolling exists, we have to ask what makes behavior *intentional*. Researchers from the field of theory of action developed many different and promising theories of intentional action. The problem is, however, that these approaches are not compatible with one another, and every approach is heavily debated. We, therefore, decided not to select one of the existing theories of intentional action, but instead rely on *minimal conditions* for intentional action. We think that if an action does not fulfill the minimal conditions, it cannot count as an intentional action regardless of the theory of intentional action one has in mind.

We consider a *necessary* (but not sufficient) condition for an action to be intentional that there is a represented personal-level goal, and that the goal causes an adequate effect (Moors & De Houwer, 2006, p. 306). We propose the following as minimal conditions for intentional action: the *action can be intentional when (a) a represented goal is present; (b) the content of the represented goal is actually realized; and (c) this occurs through a causal link between (a) and (b)*. At the description level, the goal should not contain the action’s desired result but the action as such. For instance, throwing a dart is an adequate description for the goal of throwing an arrow on the dartboard. The relevant effect is the actual action. If the action is intentional, then the action must be identical to the content of the represented goal. As the last condition, the counterfactual assumption must apply: if the goal had not been represented, the effect would not have occurred.

First (a), regardless of which mental states are constitutive of intentional actions, the intention’s content must represent the resulting action. For this purpose, it does not matter whether intentions are attitudes toward plans (Mele & Moser, 1994), components of plans (Bratman, 1984), or goals (Amaya, 2018). Intentional action does not necessarily presuppose the intention to execute this very action (e.g., like intentionally *A*-ing presupposes the intention to *A*), but at least a reasonable connection. Accordingly, a tennis player can intentionally hold a tennis racket even though they only have the intention to play tennis.

Regarding the second criterion (b), we do not want to claim that every intentional action has to be successful. Still, intentional action requires success in executing an intention. For instance, when playing darts the behavioral pattern of throwing a dart shows the execution of the intention. For this case, it is irrelevant that the subject missed the bull’s eye. We cannot give clear-cut criteria for when intentions are executed. However, we think we can rely on a purely intuitive approach as long as no empirical data or arguments speak against the intuitive interpretation of an individual case.

Finally, the last criterion (c) highlights the causal connection between the intention and the action behavior. We suggest that a counter-factual relation between the represented goal and the resulting behavior is necessary for intentional action.

When actions do not match with these three minimum criteria (a-c), the behavior is not intentional, but can still be *goal-directed*. Every behavior which serves some kind of goal is at least goal-directed, even though it is not an intentional action. In the following, when we speak of “goal-directed behavior,” we refer to behavior that does *not* fulfill all three criteria.

**2.2 A trolling example**

In the Bavarian village of Traunstein, an asylum seeker allegedly raped a girl in an underpass. As this news spread like wildfire on the social networks, the police investigated the case. It turned out, there was no sign of rape. In addition, the police identified the source of the false report. The man claimed he received the information from a secure source; however, the story turned out to be pure hearsay (Haupt, 2016). Some days later, a Bavarian broadcasting channel identified the source of this false report and confronted the person with the facts. The person acknowledged that there was no evidence for the story. At the same time, however – and this is remarkable – the person stated that the information could *still* be true. In fact, the conversation indicated that the person still thought the story was worth defending. The person suggested the police may not be allowed to report the event, or that the injured victim had not filed a complaint (Bartlewski & Henkenhaf, 2016).

We think it is sufficient for trolling to spread false reports combined with a resistance to facts and as a form of behavior that occurs repeatedly in social networks. The behavioral pattern appears to be "repetitive, intentional and harmful actions" (Shachaf & Hara, 2010, p. 363). It can be mistaken for an act of *intentional* trolling because the actors spread and defend false news despite not having any evidence. But, as mentioned in the introduction, the common *psychological* definition of trolling behavior does not focus on intentions. Accordingly, there is a chance that the psychological definition is broad enough to include cases of unintentional trolling as well. However, to show that unintentional trolling fits the psychological definition, we first have to show in which way trolling can be unintentional.

**2.3 Unintentional but goal-directed behavior**

We analyze these behavioral trolling patterns as unintentional but goal-directed behavior. We maintain that unintentional goal-directed behaviors are often the result of learned heuristics. Consider the following example: Bob and Sally have been in a romantic relationship for years. Especially in the beginning of their relationship they had a lot of disputes and controversies over different topics. At some point, Bob’s and Sally’s behavior changed as if they adapted their behavior to each other: There are hardly any more disputes between them. What happened? Sally had no intention to do so, but she used the tit-for-tat strategy: “(1) Cooperate ﬁrst, (2) keep a memory of size one, and (3) imitate your partner’s last behavior” (Gigerenzer, 2007, p. 62). Sally always tried to be cooperative, but if Bob acted uncooperatively, she acted uncooperatively as well. This behavioral pattern just turned out to work over the years, and Sally does not even recognize this pattern. However, without being aware of it, Bob adjusted his behavior to Sally’s tit-for-tat strategy. Ever since, their relationship has improved.

According to Gigerenzer (2007), a lot of our social behavior is not based on deliberative reasoning, but on *social heuristics*. The cognitive demands of navigating the social environment are very high. Given a certain social situation, a subject has to immediately interpret a flood of information. Social heuristics are very important, because we do not have the time or the cognitive resources to reflect on all the available information. Research in social psychology has revealed different heuristics, which are used to successfully navigate through the social world. Heuristics are usually paraphrased like in the example mentioned above, but this does not mean that a subject has an explicit set of rules in their mind. On the contrary, psychologists empirically research heuristics because subjects usually are unaware of them. However, subjects might be aware of the heuristics’ results and refer to them as their gut-feeling.

Heuristics guide social behavior that is often goal-directed, but unintentional. The social behavior is unintentional as there is no adequate *goal-representation*, like “use tit-for-tat strategy.” Even though there might be a goal-representation to improve the relationship, the conscious plan does not include the tit-for-tat strategy. Nevertheless, the behavior is goal-directed: The goal is to reduce controversies with Bob. Due to trial-and-error and recognizing certain environmental circumstances, Sally *automatically* developed a suitable strategy. In this sense, a lot of our social behavior happens to be conscious, but *unintentional* goal-directed behavior.

Nisbett and Wilson (1977) questioned the assumption that subjects always know what is happening in their minds. The authors compiled a variety of studies, and concluded that subjects often cannot recognize which effects a particular stimulus has on cognitive processes, like conscious judgements.[[5]](#footnote-6) Building on this, Wilson (2004) later coined the term “adaptive unconscious.” The so-called adaptive unconscious has the power to unconsciously set and change subjects’ goals. Wilson illustrates this with a short narrative. An adult and a child play tennis, and the adult has the goal to let the child win. However, while playing tennis, the adult suddenly notices that the child has lost all desire to play due to frustration. Because of that, the adult realizes that they are completely exhausted – the adult recognizes that they played to win. One possible psychological explanation is that the situational circumstances led to an unconscious change of the leading goal-representation. Without noticing it, the adult executed the usual behavioral patterns for being on a tennis court.

Likewise, we take the illustrated example of trolling (section 2.2) as an instance of *unintentional goal-directed behavior*. We hold that this behavior is unintentional, since the actor has the personal-level goal to use language descriptively, but actually uses it expressively: they articulate their affections. Our claim is as follows:

1. Acts of speech are intentional actions.
2. A behavior is intentional when (a) a represented goal is present; (b) the content of the represented goal is actually executed; and (c) this occurs through a causal link between (a) and (b).
3. There are cases in which the following applies: (a) speakers have the represented goal of using language descriptively; (b) language behavior occurs (c) on the basis of the represented goal, whereas (b) the language behavior is *not* descriptive but an expressive use of language.

The type of trolling we are analyzing is found in (3). In order to argue for (3)-cases we will show in section 3 that utterances can be expressive, even though they look descriptive.

We suggest that people often represent the goal of descriptively using language, and that this also has a counterfactual causal relationship to speech behavior. However, some cases of speech behavior, like when unintentional trolling happens, are *not* identical with the represented goal of using language descriptively, but with expressing emotions. This speech behavior is neither part of some plan (Bratman, 1984; Mele & Moser, 1994) nor is it a means to achieve a goal (Amaya, 2018). In this sense, when agents express emotions under such circumstances, they are not acting intentionally. As explained, the adaptive unconscious (Wilson, 2004) either unconsciously sets the represented goal or changes it. For these cases, the *situational* *conditions* (more on this in section 6) are important because they can trigger certain goals that can cause the behavior in question.

If we look at (unintentional) trolling from this perspective we can understand false reports differently. Sometimes people do not intentionally distribute false information to stir up prejudices, although this is often the consequence. We think that many false reports serve an expressive function for agents because they enable them to communicate their emotions. New technologies, such as social media, allow people to express prejudices through language behavior that seems descriptive but is, indeed, expressive. Referring to our example, we can explain the growing number of false reports about asylum seekers in Germany (and elsewhere) without having to rely on malicious intent as an explanation. Certainly, there are cases in which people deliberately and strategically spread false reports, but generalizing these individual cases to *all* cases is misleading.

Often, when people spread false reports they exhibit a resistance to facts, so that new information cannot disprove previous beliefs. Perceivers assume that the uttered beliefs are descriptive information with a truth value. If, however, we take such utterances as expressive, the situation changes, since expressive statements do not have a truth value. Because these utterances are just expressions of emotions, it is easy for subjects to come up with new expressions if former ones were criticized. As an upshot of this explanation, we will show that it is not necessary to classify this use of language as irrational or malicious.

**3. What appears descriptive can take on an expressive function**

In this section, we argue that utterances that appear descriptive can actually take on an expressive function. We base our argument on the fact that the theory of modern racism presupposes the same descriptive-expressive dichotomy that we assume for unintentional trolling. Accordingly, we rely on findings of empirical social psychology and not on a formal language analysis.

So-called “attitudes” explain and predict many aspects of human behavior in empirical psychology (Cooper et al., 2016). An attitude is basically the overall evaluation of an object. One has attitudes towards concrete objects, like a car, or abstract objects, like social groups. An attitude consists of three components: an affective, a cognitive and a behavioral component. We focus on the affective component, which is of special relevance to prejudice research. In social psychology, a negative affective attitude towards a social group is a *prejudice*. Psychologists usually use questionnaires to determine explicit attitudes towards objects.[[6]](#footnote-7) To directly measure explicit attitudes with a questionnaire, psychologists assume that the subject can access the mental state in question and that the subject is willing and able to adequately describe their mental state.

The justification-suppression model (JSM; Crandall & Eshleman, 2003) is a framework that explains empirically supported theories regarding the relation between expressing and suppressing prejudices. Roughly speaking, JSM is an empirical theory about expressing prejudice and the factors that lead to a high or low expression of prejudice. According to the JSM, the expression of prejudice results from the subject’s genuine prejudice and the suppression factors (“prejudice + suppression = expression,” Crandall & Eshleman, 2003, p. 416).

According to JSM, a genuine prejudice is a mental state with affective content that *motivates* behavior. However, genuine prejudice is neither identical to *implicit* nor equivalent to *explicit* attitudes. On the contrary, these measured (implicit/explicit) attitudes are already what JSM calls “expressions.” These expressions are the result of genuine prejudices that are tunneled through suppression and justification processes. Suppression “is an externally or internally motivated attempt to reduce the expression or awareness of prejudice” (Crandall & Eshleman, 2003, p. 420). *Social norms* are one of the most important factors for the suppression of prejudices. Western societies are – by and large – committed to egalitarian values and sanction prejudice through social norms. This explains why the willingness to express prejudices has been decreasing for decades (Bobo et al., 2012). Furthermore, in the JSM, justifications “serve as an opportunity to express genuine prejudice without suffering external or internal sanction” (Crandall & Eshleman, 2003, p. 456). Such justifications are apparent, for example, when someone claims that a person’s social standing only depends upon their decisions and individual actions (that is, the unwavering belief in meritocracy). Therefore, someone holding this view might think that minorities facing discrimination rightly suffer because they made a lot of wrong choices and that affirmative action is a form of reverse racism.

However, the JSM’s claim that justifications themselves are vehicles for expressing prejudice is heavily debated and controversial. Still, we maintain that the psychological theory of modern racism supports the JSM. The Modern Racism Scale (MRS; McConahay et al., 1981, 1986) implicitly supports the JSM from a theoretical point of view. The MRS measures prejudices because it presupposes that (a) self-reported prejudices are *suppressed* due to social norms while (b) justifications serve to *express* prejudices.

The MRS is still frequently used today (for instance, see Blatz & Ross, 2009; for an overview, see Morrison & Kiss, 2017) to measure explicit prejudices and predict discriminatory behavior (see experiment four in McConahay et al., 1986). The MRS assumes that emotions essentially guide convictions, even though agents do not want to reveal how strong their prejudices really are. Therefore, the MRS uses special items to measure people's attitudes, namely ambiguous questions. A participant can answer these questions using either racist or non-racist justifications. When measuring prejudice with the MRS, subjects have to express their agreement with the scale’s items. For example, an ambiguous item could be the question whether positive discrimination, that is, recruitment quotas, can be justified for people of color. On the one hand, subjects can answer this question negatively if they highly value personal responsibility and performance: only the best should get a job. On the other hand, subjects can also answer the question negatively on the basis of racist beliefs. Since the subject can justify their answer to themselves and others in a non-racist way, they express their emotions.

However, the MRS only measures how much participants agree or disagree with a certain statement. A possible interpretation of what happens when a participant fills out the MRS is that the participant’s *beliefs* are measured. However, at the same time, the MRS is actually used to measure explicit prejudices that are negative *emotions* towards a social group. So, does the MRS measure an attitude’s affective component or the cognitive component?

While the MRS measures beliefs, these beliefs serve to operationalize the attitude’s affective component. The MRS presupposes that subjects express their affect through belief states. The authors of the MRS state that: "... the items in the Modern Racism Scale permit the *expression* of negative affect because giving the prejudiced response in each instance can be explained by racially neutral ideology or non-prejudiced race-relevant attributions" (McConahay et al., 1986, p. 100, our emphasis). The subjects should, unbeknownst to them, express their negative emotions using the items. Psychologists infer the participant’s underlying affect by operationalizing the participant’s beliefs.

It is important to note that the subjects undergoing a prejudice measurement with the MRS do not know what the researchers are looking for. Therefore, the subjects experience their use of language as descriptive, but in fact they *express* negative emotions towards a social group. The expression of explicit prejudice is unintentional because it does not meet the minimal conditions for intentional actions: While the subjects represent the goal of using descriptive language, they express their negative racial attitudes. Accordingly, one of the minimal conditions for intentional action, namely the second (b), is not satisfied. The subject has a certain goal representation, *A*, but shows inappropriate *B*-behavior. So, if we would ask the subject, “why do you express your prejudices,” they would probably answer: “I was not aware that I was doing so.” According to Anscombe (1957), this is a clear-cut case of unintentional behavior.

Social norms force subjects to suppress their prejudices due to external social pressure. The theory of group socialization (Levine & Moreland, 1994) holds that subjects that do not conform to group norms receive admonitions. Social exclusion is a constant threat for those that are not in line with social norms. So-called *political correctness*, “an implicit social convention of restraint on public expression, operating within a given community” (Loury, 1994, p. 430), can be understood as such a social norm. We can expect that prejudiced subjects are aware of these mechanisms and suppress their prejudices.

However, subjects find ways to express their prejudices and bypass social control. Like the JSM suggests, they use justification strategies to express their prejudices without being liable to blame. In the next section, we will focus on the use of false reports and further elaborate our claim that these false reports are actually an expression of mental states. We explain this by referring to the adaptive unconscious and the JSM.

**4. You don't say things like that!**

We assume that the trolling behavior in question is unintentional (for example, there is no intention to harm or annoy), but goal-directed. In addition, the language behavior itself is unintentional as well: the intention to use language descriptively leads, in fact, to expressive language behavior. We think of these heuristically evolved behavioral patterns as the result of trial-and-error. In this section, we want to suggest the goal towards which the behavior is directed. We also analyze one way of expressing prejudices, namely spreading false reports, as a way of storytelling.

From early childhood, people learn how to behave and how not to behave. Social groups establish normative standards that all members should meet. Such social norms can include a variety of areas, such as clothing style, musical taste, or even a certain style of speech. As rules, these social norms regulate all kinds of behavior (Brennan et al., 2016). Norms also play an important role in political discourse. Ultimately, the political discourse of a society determines the political orientation of a democracy. In democracies, all eligible voters should be able to participate in this discourse. All individuals, regardless of their social class or education, should be able to participate in the same way. However, the political discourse has some limitations for participation in terms of content and form.

First, some *content* is forbidden in the political discourse due to social norms (Harell, 2010). For example, saying that homosexuals are disgusting is not considered a useful contribution to the political discourse. Statements like these are discriminatory acts, and since they only express an attitude, they are hardly a useful contribution.

It might be objected that racist speech is currently on the rise, so there is no need to express prejudices via justification-strategies, as suggested by the JSM. We think that this impression is false. Recent GSS data from 2018 (General Social Surveys, 1972-2018) show that the disapproval of racist speech has increased over the years and never was as high as it is currently. From 2014 to 2018, the number of those that thought a racist professor should not teach at a university increased. Likewise, there was more agreement to remove a racist book from the local library. Furthermore, the data shows that there was an increase of Americans that assume that inequalities between Blacks and Whites are caused by discrimination and are not self-inflicted. The German General Social Survey from 2018 indicates that Germany has a problem with racism, but that there was no substantial increase of hostility towards foreigners over the last years (Blohm & Wasmer, 2018, p. 411). In sum, we think that these statistics, measured via personal questioning, indicate that there are restrictions regarding racist speech (restriction of content). The Black Lives Matter movement and its global support by politicians, media and companies indicates a disapproval of racist speech as well.

Second, regarding *formal* restrictions of the public discourse, the ability to formulate and evaluate arguments is not equally distributed in society. This asymmetry between people with high and low levels of education (and rhetorical proficiency) leads to more political influence by those with higher levels of education (Nie et al., 1996).

Accordingly, social norms exclude some contributions to political discourse from actors because of their contribution’s content (Harell, 2010) or form (Nie et al., 1996). There are regulations regarding content (like clearly affective statements) and formal conditions (like rules of argumentation) for participation in the public discourse. Nevertheless, people find ways to contribute by *apparently* using the same methods: arguments and facts.

In this context, it is useful to remember that there are different concepts of rationality (for example, see Nozick, 1994). A formal concept of rationality typically dominates the philosophical discussion: rationality relies on strict mathematical methods, like utility theory and logic (Gigerenzer, 1991). However, this is only one facet of rationality.

Gigerenzer (1991) claims that we often call behaviors irrational because we do not consider the concrete *context-conditions* in which a subject behaves: “A small town in Wales has a village idiot. He once was offered the choice between a pound and a shilling, and he took the shilling. People came from everywhere to witness this phenomenon. They repeatedly offered him a choice between a pound and a shilling. He always took the shilling” (Gigerenzer, 1991, p. 103). *Utility theory* is one way to identify choices as rational or irrational. According to utility theory, the most valuable option for an actor is calculated, given that all choices have an uncertain outcome. The most valuable option is the *rational* option. In this framework, the behavior of the “village idiot” is irrational because he rejects the greater utility that has a constant probability of success. From this point of view, he makes a bad and thus irrational decision. However, the *context* of the actor allows him to repeat the decision as often as he wishes. Under these environmental circumstances his behavior appears to be rational.

Accordingly, if one takes into account the circumstances in which subjects find themselves, many alleged irrationalities disappear. For the unintentional troll, the context is the political discourse with its form and content restrictions. Regarding the political discourse, the agent is, so to speak, in a disadvantaged position due to external conditions. False reports or a resistance to facts may seem irrational, depending on the perspective. This classification, however, neglects the circumstances in which the actors find themselves. The actors circumvent the political discourse’s restrictions using an inventive form of communication that involves *storytelling*. By telling a story, an agent can express prejudices and avoid social sanctions.

Storytelling is a heavily discussed topic in psychology, and has played an important role in the history of psychology.[[7]](#footnote-8) From a psychological perspective, storytelling means that actors create a story that reveals something about their psychological setup. Classic measures, like the Thematic Apperception Test (TAT; for an overview, see Cramer, 2005), assume that storytelling – in the broadest sense – unintentionally reveals a person’s motives, desires, intentions or stereotypes. Analogously, Human Figure Drawings (Handler, 1996) are discussed as an instrument that uncovers participants’ mental states without their awareness.

For instance, one Human Figure Drawing task is to draw a scientist. We consider this to be *storytelling* because the created figure usually is located somewhere (like in an office or a laboratory) and performs an action (like reading a book or holding a test-tube). The task’s underlying idea is that drawing a scientist expresses the participant’s stereotype of a scientist, including properties like gender. A meta-study (Miller et al., 2018) shows that children in the US nowadays more frequently draw female scientist than in the past. A possible explanation is that the media are increasingly showing female scientists, and thereby stereotypes about scientists are slowly changing.

Fake news, as illustrated above in section 2.2, are cases of storytelling. Subjects use stories to express prejudices by apparently contributing to the political discourse in an accepted way. While people are telling stories, they apparently describe facts, which is widely considered as a valid contribution to the political discourse. The JSM suggests that people use this strategy as a *justification* for expressing prejudices while bypassing external pressure. Anything that serves the purpose of expressing prejudice while avoiding social sanctions can serve as a justification. Coming back to our example (section 2.2), the person expresses negative stereotypes and prejudices towards refugees using storytelling as a means for justification. These unintentional expressions of prejudices are produced through a communicative strategy that evolved heuristically. This explains why people in such situations reject critical feedback because expressions of emotions are neither true nor false – nor are they justified in any relevant sense.

The criteria we have presented for unintentional trolling are consistent with the psychological definition of trolling: “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 et al., 2014, p. 97). Even if trolling is unintentional, we have shown that this behavior is actually deceptive, destructive and disruptive. Individuals that troll unintentionally are *deceptive* towards themselves and others: they do not recognize what they are doing, namely expressing emotions, while other individuals falsely perceive their use of language as descriptive. Furthermore, this explains why such an individual is resistant to counterarguments: expressions of emotions are neither true nor false. As the effects of fake news show, this resistance to facts leads to *disruptions* in social media. False reports, as expressions of prejudices, are a vehicle of negative stereotypes and negative emotions, which is why unintentional trolling can have *destructive* consequences.

**5. Distinguishing unintentional trolls from other trolls**

If our hypothesis is correct and some individuals troll unintentionally in social media networks, how many of them are there? This is an empirical question that we cannot answer. However, although we cannot look into a another person’s mind, we can expect different patterns of behavior for classic trolling than for unintentional trolling. In this section, we propose distinctive behavioral patterns of unintentional trolling that serve to distinguish it from classical trolling. As a result, these criteria could also be used, in principle, to detect and to estimate the number of unintentional trolls in social media networks.

So, what distinguishes unintentional trolls from classic ones?

* They do not use fake accounts

Since people that troll unintentionally do not want to harm or annoy, they do not hide behind a fake account. In contrast, using fake accounts is very common for classic trolling (Phillips, 2016), cyber-bullying (Guo, 2016), and cyber-harassment (Li & Beran, 2016).

* They tell stories to express their attitudes

If subjects want to express their prejudices and, at the same time, do not want to be socially sanctioned, then they need a justification. We suggest that storytelling is a very efficient way to express prejudices while being perceived as non-prejudiced. As a good story is hard to verify as false, it is a useful strategy to be as vague as possible regarding the information’s source. Therefore, if unintentional trolling happens, the subject will obscure the source of information (“I know somebody who knows somebody…”). A different way to tell a story which serves as a justification is to insist that the storyteller saw the event themselves. This is a good strategy because it makes the information unverifiable for others.

* They do not use memetic content

In contrast to classic trolls, who are aiming for *lulz,* unintentional trolls do not refer to memetic content. As described above, trolls use memes to increase *lulz*. Even though unintentional trolling may use stereotypes and narratives, which vary between racism, sexism or antisemitism, these narratives are not memes by themselves. In the current context, memes are not just units of culture, as Dawkins (1989) suggested, but are half-baked jokes that entail a certain history of use.

* They do not do it for the *lulz*

A part of unintentional trolling consists of resisting arguments that would disprove the subject’s expressed affect. However, because expressive utterances are never true nor false, arguments are ineffective. Nevertheless, such trolls, even though their behavior might be misleading, do not have the goal to make fun of others’ outrage. In contrast, when a classic troll successfully trolls somebody, we can expect an inappropriate, infantile and counterproductive style of argumentation that tries to create more rage.

**6. Conclusion**

It is not our aim to challenge classical trolling theories but to complement them. Unintentional trolling is usually also repetitive and harmful, but it lacks the intentionality that classical trolling entails. We take unintentional trolling to be a goal-directed behavior. Therefore, there is no goal-representation that causes adequate behavior. In contrast, the behavior results from heuristically learned adaptions that serve to achieve a certain goal, namely to circumvent social norms of public discourse. Using the justification-suppression model, we argued that telling made up stories is an efficient way to express prejudices and stereotypes. Storytelling, in psychology, is traditionally assumed to reveal something about a person’s psychological setup.

The motivation for unintentional trolling is being able to participate in political discourse. The political discourse, however, has limitations to participation regarding form and content. Regarding formal restrictions, standards of argumentation are more demanding for some individuals than for others. Thus, an asymmetry arises due to educational disparities. Regarding content, some claims are generally not accepted, since social norms regulate the discourse. Nevertheless, some agents want to bring their views into the discourse or see them discussed. Considering the circumstances in which the agents find themselves, their strategy is only irrational if one neglects the environmental context of their behavior.

**7. References**

Amaya, S. (2018). Two kinds of intentions: A new defense of the simple view. *Philosophical Studies*, *175*(7), 1767–1786. https://doi.org/10.1007/s11098-017-0934-1

Anscombe, G. E. M. (1957). *Intention*. Blackwell.

Arizona State Legislature. (2012, April). *HB2549—502R - I Ver*. <https://www.azleg.gov/legtext/50leg/2r/bills/hb2549p.htm>

Austin, J. L. (1962). *How to do things with words* (J. O. Urmson & M. Sbisá, Eds.; 2nd Edition). Harvard University Press.

Bartlewski, M., & Henkenhaf, T. (2016, April 30). Wie gefährlich sind Fakes im Netz? [How dangerious are fake-news?]. *Bayrischer-Rundfunk*. <http://blog.br.de/die-frage/fake/>

Blatz, C. W., & Ross, M. (2009). Principled ideology or racism: Why do modern racists oppose race-based social justice programs? *Journal of Experimental Social Psychology*, *45*(1), 258–261. <https://doi.org/10.1016/j.jesp.2008.08.008>

Blohm, M., & Wasmer, M. (2018). Einstellungen und Kontake zu Ausländern. In Statistisches Bundesamt & Wissenschaftszentrum Berlin für Sozialforschung (Ed.), *Datenreport 2018—Ein Sozialbericht für die Bundesrepublik Deutschland [Datareport 2018—A social report for the Federal Republic of Germany]* (pp. 403–411). Bundeszentrale für politische Bildung.

Bobo, L. D., Charles, C. Z., Krysan, M., & Simmons, A. D. (2012). The real record on racial attitudes. In P. V. Marsden (Ed.), *Social Trends in American Life: Finds from the General Social Survey since 1972*. Princeton University Press.

Bratman, M. (1984). Two faces of intention. *Philosophical Review*, *93*(3), 375–405. <https://doi.org/10.2307/2184542>

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2016). *Explaining norms* (Reprint). Oxford University Press.Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Cooper, J., Blackman, S., & Keller, K. (2016). *The science of attitudes* (1st Edition). Routledge.

Cramer, P. (2005). *Storytelling, narrative, and the Thematic Apperception Test* (1st Edition). Guilford Publications, Inc.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Dawkins, R. (1989). *The Selfish Gene* (2nd Revised Edition). Oxford Paperbacks.

Fisher, W. R. (1985). The narrative paradigm: An elaboration. *Communication Monographs*, *52*(4), 347–367. https://doi.org/10.1080/03637758509376117

Gawronski, B., & Payne, P. (2010). A history of implicit social cognition. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*. Guilford Press.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious* (1st Edition). Viking Adult.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases. *In W. Stroebe & M. Hewstone (Eds.), European Review of Social Psychology*, 83–115.

Green, Melanie, & Brock, T. (2000). The role of transportation in the persuasiveness of public narrative. *Journal of Personality and Social Psychology*, *79*, 701–721. <https://doi.org/10.1037/0022-3514.79.5.701>

Green, M. (2017). Speech Acts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2017/entries/speech-acts/>

Guo, S. (2016). A meta-analysis of the predictors of cyberbullying perpetration and victimization. *Psychology in the Schools*, *53*(4), 432–453. https://doi.org/10.1002/pits.21914

Handler, L. (1996). The clinical use of drawings: Draw-A-Person, House-Tree-Person, and kinetic family drawings. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 206–293). Allyn & Bacon.

Harell, A. (2010). Political tolerance, racist speech, and the influence of social networks. *Social Science Quarterly*, *91*(3), 724–740. JSTOR.

Haupt, F. (2016, January 26). Flüchtlinge: Stimmung machen mit Gerüchten [Refugees: Turing the public opinion against refugees through rumors]. *Frankfurter Allgemeine Zeitung*. <http://www.faz.net/aktuell/politik/inland/fluechtlinge-wie-mit-geruechten-stimmung-gemacht-wird-14030797.html>

Levine, J. M., & Moreland, R. L. (1994). Group socialization: Theory and research. *European Review of Social Psychology*, *5*(1), 305–336.

Li, Q., & Beran, T. (2016). Cyber-Harassment: A study of a new method for an old behavior: *Journal of Educational Computing Research*, *32*(3), 265–277. https://doi.org/10.2190/8YQM-B04H-PG4D-BLLH

Loury, G. C. (1994). Self-Censorship in public discourse: A theory of “political correctness” and related phenomena. *Rationality and Society*, *6*(4), 428–461. https://doi.org/10.1177/1043463194006004002

McConahay, J. B., Hardee, B. B., & Batts, V. (1981). Has racism declined in America?: It depends on who is asking and what is asked. *Journal of Conflict Resolution*, *25*(4), 563–579.

McConahay, J. B., Hardee, B. B., & Batts, V. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

Mele, A. R., & Moser, P. K. (1994). Intentional action. *Noûs*, *28*(1), 39–68. https://doi.org/10.2307/2215919

Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children’s gender-science stereotypes: A meta-analysis of 5 decades of U.S. Draw-A-Scientist studies. *Child Development*, *89*(6), 1943–1955. https://doi.org/10.1111/cdev.13039

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326.

Morrison, T. G., & Kiss, M. (2017). Modern Racism Scale. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–3). Springer International Publishing. <https://doi.org/10.1007/978-3-319-28099-8_1251-1>

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *48*(3), 231–259.

Nie, N. H., Junn, J., & Stehlik-Barry, K. (1996). *Education and democratic citizenship in america*. University of Chicago Press.

Nozick, R. (1994). *The nature of rationality*. Princeton University Press.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In R. E. Petty & J. T. Cacioppo (Eds.), *Communication and persuasion: Central and peripheral routes to attitude change* (pp. 1–24). Springer. <https://doi.org/10.1007/978-1-4612-4964-1_1>

Phillips, W. (2016). *This is why we can’t have nice things: Mapping the relationship between online trolling and mainstream culture* (Reprint). The MIT Press.

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, *36*(3), 357–370.

Shifman, L. (2013). *Memes in digital culture*. MIT Press.

Smith, T. W., Davern, M., Freese, J., and Morgan, S., *General Social Surveys, 1972-2018*. Principal Investigator, Smith, T. W.; Co-Principal Investigators, Davern, M., Freese, J., and Morgan, S.; Sponsored by National Science Foundation. Chicago: NORC, 2018. Data accessed from the GSS Data Explorer website at [gssdataexplorer.norc.org](http://gssdataexplorer.norc.org).

Wilson, T. D. (2004). *Strangers to ourselves: Discovering the adaptive unconscious* (1st Edition). Belknap Press.

1. When we talk about memes, we are not thinking about memes in the sense that Dawkins (1989) suggested. In contrast, we stick to the characterization by Shifman, who describes memes as “(a) a group of digital items sharing common characteristics of content, form, and/or stance; (b) that were created with awareness of each other; and (c) were circulated, imitated, and transformed via the Internet by many users.” (Shifman, 2013, p. 6) [↑](#footnote-ref-2)
2. The *Pew Center on Internet and Technology* is a great online resource about trolling and related topics. [↑](#footnote-ref-3)
3. Sometimes, fake news is defined as intentionally spread false news in order to manipulate the receivers. However, we think that this definition is question begging – we suggest that some false-news are spread by actors without these harmful intentions. [↑](#footnote-ref-4)
4. Whether this behavioral pattern can count as trolling is debatable. We think, from the third person perspective the observed behavior appears trolling-like. In 2.2 we give an example which makes our point more clearly. [↑](#footnote-ref-5)
5. The whole modern *implicit* social psychology is based on this assumption (see, for example, Gawronski & Payne, 2010). [↑](#footnote-ref-6)
6. In this paper, we only focus on explicit attitudes. [↑](#footnote-ref-7)
7. Walter Fisher presents the Narrative Paradigm Approach (NPA; Fisher, 1985) as a philosophical theory of communication dealing with the crucial value of stories for human beings. According to Fisher, humans are essentially storytellers. Because within the NPA communication grounds on stories, there is no sharp distinction between *arguments* and *stories*. However, we think that this distinction is crucial for analytic philosophy. Furthermore, sophisticated theories of attitude persuasion (for example, Petty & Cacioppo 1986) show that it *does* make a substantial difference if a subject evaluates anargument as opposed to a story. In fact, Green and Brock (2000) investigated the underlying mechanisms for attitude persuasion through stories. For these cases, they suggest that the receiver’s salient cognitive mechanisms are substantially distinct to those that are responsible for attitude change via arguments. Accordingly, we think that from the stance of behavioral science it is justified to assume that there is a difference between stories and arguments. [↑](#footnote-ref-8)