

# Bullshit, Pragmatic Deception, and Natural Language Processing

**Oliver Deck**

*Ruhr University Bochum*

OLIVER.DECK@RUB.DE

**Editor:** Vera Demberg

Submitted 06/2022; Accepted 05/2023; Published online 05/2023

## Abstract

Fact checking and fake news detection has garnered increasing interest within the natural language processing (NLP) community in recent years, yet other aspects of misinformation remain unexplored. One such phenomenon is ‘bullshit’, which different disciplines have tried to define since it first entered academic discussion nearly four decades ago. Fact checking bullshitters is useless, because factual reality typically plays no part in their assertions: Where liars deceive about content, bullshitters deceive about their goals. Bullshitting is misleading about language itself, which necessitates identifying the points at which pragmatic conventions are broken with deceptive intent. This paper aims to introduce bullshitology into the field of NLP by tying it to questions in a QUD-based definition, providing two approaches to bullshit annotation, and finally outlining which combinations of NLP methods will be helpful to classify which kinds of linguistic bullshit.

**Keywords:** bullshit, NLP, pragmatics, misinformation, QUD

## 1. Introduction

Common parlance applies the term *bullshit* to nonsensical utterances, lies, personal affronts, injustices, and any number of daily annoyances. But from a linguistic perspective, bullshit presents itself as a highly complex, amorphous pragmatic concept which calls for novel methods when approached from a natural language processing (NLP) perspective. For some bullshit statements, context is extremely important while others may reveal themselves through textual analysis alone. Consider the following statement made by former U.S. president Donald Trump:

- (1) I am the least racist person anybody is going to meet. (Trump, 2018)

It may not be immediately evident, why this statement is bullshit from a linguistic perspective. Here, the interpretation does not stem from how one views Trump. Rather, it is based on the form of the utterance: He could have chosen to lie, e.g. by saying *No one has ever accused me of racism*, which would have been an assertion that can be true or false. Instead, he chose to use the form of an assertion to make a statement about an unquantifiable and thus unverifiable concept. He in essence deceived about the pragmatics of the utterance itself. The statement can be analyzed as bullshit without the need for additional textual context since one cannot make such an assertion with implied certainty about one’s degree of racism.<sup>1</sup>

1. Although world knowledge is needed to ascertain which concepts are inherently unquantifiable. We also need to take into account the audience: A favorable listener might interpret the utterance as an hyperbolic way of saying *I am*

Other examples of bullshit require careful analysis of the larger context or of utterance-external factors: Evasive answers in interviews or exams, for example, can only be interpreted as bullshit when taking into account the questions they fail to answer. Over the past four decades, the field of ‘bullshitology’ has investigated a number of related phenomena that are generally deceptive in nature, but, crucially, cannot be classified as outright lies. Since its introduction into academia in the highly influential paper “On Bullshit” by American philosopher Harry G. Frankfurt (Frankfurt, 2005), researchers in different disciplines have tried to define what exactly bullshit is. However, while audiences may intuitively know when to call someone out on it, the scientific concept of bullshit remains unclear and many researchers have proposed their own definitions of these deceptive practices outside of the truth-untruth dichotomy.

Much computational research has, in recent years, focused on the linguistics of lying, deception, fake news, propaganda, and a host of related phenomena – but so far not on bullshit. Shared tasks on fake news detection consistently garner considerable interest, leading to improvements in automated systems (Zhou and Zafarani, 2020; Zeng et al., 2021; Shahi et al., 2021). We cannot identify bullshit via content analysis, since, unlike lies and fake news, it does not depend on being factually wrong. Rather, it is a kind of deception about language use, which calls for linguistic analysis. Frankfurt may well have exaggerated when he called bullshit a “greater enemy of the truth than lies” (Frankfurt, 2005, p. 15), but in the context of ‘alternative facts’ and ‘post-truth,’ it is imperative to approach bullshit as one subclass of disinformation.

This paper approaches bullshit for the first time from a computational linguistics perspective. By its nature, computational linguistics can rarely look at the producer – the bullshitter – and typically focuses on the product, in this case textual or spoken bullshit. We typically have no insight into the cognition, situation (except from a broader, pragmatic context), or intention of an author and must therefore base our methods on text alone, so deciding whether an utterance like (1) is bullshit (in the linguistic sense), and not lying, hyperbole, or some other phenomenon is challenging. It relies on a workable definition of bullshit, narrow enough to be applicable to very specific kinds of writing or speech, but broad enough to encompass at least those examples of bullshit on which previous research agrees. The definition must also be operationalizable, i.e. it must describe bullshit that can be classified via existing – or at least theoretically possible – methods of computational linguistics. From this need follows the first research question:

**RQ 1:** What is a useful bullshit definition for a computational linguistics approach?

In this paper, I argue that the deception inherent in bullshit is a deception about pragmatics, undermining conventional norms of discourse. Where liars typically use language to deceive about the real world, about concepts, feelings, or their state of mind, bullshitters deceive about language itself. They use language to mislead about their language use. Bullshit happens at the breaking points of pragmatic models, where cooperative discourse is disregarded. In saying *I am the least racist person anybody is going to meet*, Donald Trump neither lies nor does he tell the truth. In fact, with a statement of this kind, it is impossible to do either. The statement is presented as an assertion, but is unverifiable by its nature. Rather, the language of assertion is used to present an image of the speaker – an ‘expressive’ masquerading as an ‘assertive’ in the nomenclature of Searle (1976) – shrouding the actual intent of the speaker but crucially not lying about anything. To find

---

*not racist*. When exactly hyperbole is acceptable as a stylistic choice and when it becomes bullshit is dependant on extralinguistic factors of speaker, situation, and audience.

a sufficiently large number of bullshitting examples, we need to create processes and guidelines of bullshit annotation and analysis. The second research question is therefore:

**RQ 2:** How can we annotate and analyze bullshit?

The third research question concerns the relevant NLP methods for bullshit detection:

**RQ 3:** What are promising NLP approaches to bullshit classification?

Computational pragmatics is a complex and fairly recent field which is still far from solved in almost all areas of inquiry, yet we may connect pragmatic models to related computational approaches. While we cannot directly build on the vast amount of fact-checking research, other fields may help in stages of the bullshit detection process: claim identification (e.g. Hassan et al., 2015; Barrón-Cedeño et al., 2020), question answering (e.g. Choi et al., 2018; Kim et al., 2021), question/answer congruity (e.g. Faruqui and Das, 2018; Yu and Jiang, 2021), argument quality (e.g. Skitalinskaya et al., 2021), propaganda detection (e.g. Barrón-Cedeño et al., 2019a,b; Da San Martino et al., 2020a,b), and others. Since such issues have garnered greater attention in the NLP community, synergistic effects may be utilized for bullshit classification and detection.

The next chapter will answer **RQ 1** by looking at the history of bullshit research and formulating a definition that serves as a basis for our analysis. Chapter 3 will outline different avenues of operationalizing the bullshit annotation process in order to answer **RQ 2**, reporting on a small pilot study on annotating a subtype of bullshitting. Finally, chapter 4 will provide an outlook on the potential of natural language processing in automatically classifying bullshit in an answer to **RQ 3**.

## 2. What is Bullshit?

### 2.1 The History of Bullshitology

Frankfurt originally describes bullshit as those statements in which speakers do not care whether what they say is true or false (Frankfurt, 2005). Liars and truth-tellers must at least have some notion of the truth to either distort or affirm it; Bullshitters have different goals, such as portraying a specific self-image. This simple view of statements without caring about the truth has since been called into question and refined by a number of researchers from different fields. Many agree with Frankfurt on the dangers of bullshit, while others have studied more harmless forms and highlight a number of positive and prosocial aspects of bullshitting; e.g. in a study on hitchhikers in America (Mukerji, 1978). In these situations, where all participants expect bullshit, the practice can take the form of a language game aiding speakers' identity work (cf. Mears, 2002), fulfilling functions like socialization, expressing feelings, passing time, resolving personal or interpersonal strain, impression management, and others (Mears, 2002). Others do not classify such playful forms as bullshit, especially when all participants are aware of the practice and there is no deception involved (e.g. Stokke and Fallis, 2017).

On larger, societal scales, bullshit can become harmful by weakening the social and linguistic norms of what constitutes acceptable, cooperative discourse. Spicer argues that, in workplace settings, harmless bullshit can become harmful when performing the form becomes more important than the content (cf. Spicer, 2020). Similarly, so-called pseudo-profound bullshit<sup>2</sup> as in (2) may

---

2. Pseudo-profound bullshit comprises non-profound, inane statements masquerading as profound. Note, neither conventionally profound statements such as "A wet person does not fear the rain" (Pennycook et al., 2015, p.549) nor

seem humorous, but research has shown that individuals who ascribe meaning to such sentences are also prone to believe in alternative medicine, angels, and pseudoscience, as well as fake news or propaganda (Pennycook et al., 2015; Pennycook and Rand, 2018, 2020, 2021; Littrell et al., 2021a,b). Believing some kinds of bullshit may thus not in itself be harmful to society, but can be an indicator for ways of thinking that may be susceptible to exploitation by bad-faith actors.

(2) Wholeness quiets infinite phenomena (Pennycook and Rand, 2020, p. 196)

Following Frankfurt’s view of the dangers of bullshit itself, a number of researchers highlight the need to deal with it similarly to lying and general disinformation. Focusing on the product instead of the producers, Cohen notes that ‘academic bullshit’ is harmful to academic practice (Cohen, 2002), while others seek to educate the public about bullshit because of its negative consequences both on individuals and society as a whole (Sagan, 1996; Bergstrom and West, 2020; Petrocelli, 2021).

In a rejection of Frankfurt’s simple definition of bullshit, Carson (2016) first introduced the distinction between persuasive and evasive bullshit, which has since brought to light striking differences in the situations and reasons for bullshitting, as well as in bullshitters’ perception and cognitive abilities. Whereas Frankfurt dealt with speakers using persuasive bullshit propositions to present themselves (often unprompted) in a specific light, evasive bullshit tends to occur when prompted: Politicians may employ evasive bullshitting to avoid giving a straight-up answer; students may not know the answer to an exam question but hope that simply saying something might award them some points as in (3). In such cases, the speaker may very well care about the truth value of what they are saying, yet try to hide that they are not really answering the question. The split between persuasive and evasive bullshit has since been substantiated by experimental psychology, showing measurable cognitive differences between those who more often persuasively bullshit and those that use bullshit as an evasive tactic (Littrell et al., 2021a,b).

(3) **Careful Test Taker**

A student who gives a bullshit answer to a question in an exam might be concerned with the truth of what she says. Suppose that she knows that the teacher will bend over backwards to give her partial credit if she thinks that she may have misunderstood the question, but she also knows that if the things she writes are false she will be marked down. In that case, she will be very careful to write only things that are true and accurate, although she knows that what she writes is not an answer to the question. (Carson, 2016, p. 62)

In linguistic research outside of bullshitology, there is also a rich history of inquiry into evasive speech. Greatbatch investigates ‘agenda-shifting behaviors’ in political interviews, where interviewees shift the topics of questions and provide answers to more favorable (to them) subjects. This idea is similar to the one described in Section 3.2 based on Gabrielsen et al. (2020). Political interviews also serve as the basis for a Bull and Mayer (1993) paper on avoiding questions. The authors develop a typology of 11 avoidance categories with 30 subcategories, most of which cannot be classified as bullshitting behavior. Among them are ignoring the question, questioning the question, attacking the question, attacking the interviewer, and others. One category that overlaps with evasive bullshitting is ‘Makes political point’ which includes presenting or justifying policy, appeals to

---

pseudo-profound statements as in (2) must be factually true or false. Instead the latter deceives about being the same linguistic device as the former.

nationalism, and self-justification (Bull and Mayer, 1993, p. 656-661). As will be shown in Section 3.2, such shifting maneuvers can be interpreted as one type of bullshitting.

Łupkowski and Ginzburg (2016) utilize corpora to look at questions that are in turn answered by questions on a larger scale. Ginzburg et al. extend the model to a “formally underpinned characterization of the response space of questions.” (Ginzburg et al., 2022, p. 40) Their typology consists of three categories of answers: question-specific, clarification responses and evasion responses. Question-specific responses are either answers or dependent questions (Ginzburg et al., 2022, p. 2). Evasion responses can state that it is difficult to answer, question the motive of the original query, or change the topic (cf. Bull and Mayer, 1993). Queries can also be evaded by ‘ignoring,’ which in the context of this model describes “address[ing] the situation, but not the question” (Ginzburg et al., 2022, p. 2). This behavior could be classified as bullshitting as described in Section 3.2.

While evasive speech has thus prompted linguistic analysis for quite some time, when it comes specifically to *bullshit*, there are only two major definitions: Meibauer (2016, 2020) and Stokke and Fallis (2017). Meibauer bases his definition of bullshit on four factors: assertion, loose concern for the truth, misrepresentational intent, and too much certainty (Meibauer, 2016, p. 75). In his view, bullshitters do not mislead about the content of their assertion (as liars do), but instead about their own loose concern for the truth of these assertions. Meibauer’s definition is a good starting point for purely linguistic inquiry but it is unclear if and how the four factors may be operationalized computationally: While we may find linguistic markers for certainty, the loose concern for the truth and misrepresentational intent may be impossible to ascertain with purely linguistic means.

From an NLP standpoint, Stokke and Fallis’ approach could therefore prove more fruitful. The researchers look at bullshit through the lens of Questions under Discussion (QUD, see Roberts, 2012), subsuming Frankfurt’s original indifference-to-truth definition of bullshitting (Stokke and Fallis, 2017, p. 279). Where Frankfurt assumed the bullshitter to be indifferent to the truth at a content level, Stokke and Fallis propose a definition in which speakers are indifferent to whether or not their utterance constitutes a truthful answer to a QUD. The approach thus covers cases like bullshitting while caring about the truth – as in (3) – and bullshitting while lying.<sup>3</sup> As noted earlier, Stokke and Fallis disregard most prosocial, playful kinds of bullshitting, since they do not consider these utterances real assertions occurring in serious situations (Stokke and Fallis, 2017, p. 290).

The two linguistic theories again highlight that bullshit is deceptive on a different level than lying. For Meibauer, bullshitters deceive about intention, for Stokke and Fallis, they deceive about their cooperative participation in QUD-based discourse. Especially the latter definition hints at bullshitting being a phenomenon that breaks pragmatic conventions. Linguistic analysis points toward a disregard not for the content, as originally postulated by Frankfurt, but rather for the linguistic practices and norms themselves. Stokke and Fallis’ definition does not encapsulate all kinds of especially evasive bullshitting, but it can serve as a starting point for NLP methods, since the QUD approach ties into computational fields related to question answering, question-answer-congruence and others.

Especially with the release of ChatGPT (OpenAI, 2022), a number of news outlets, blogs, and AI researchers have been calling large language models (LLMs) bullshitters (e.g. Bernoff, 2022; Narayanan and Kapoor, 2022; Nast, 2022). There is as of yet no formal analysis of LLMs as bullshitters though some writers refer to the Frankfurtian definition. However, following the view

---

3. Carson (2016, p. 60f.) gives an example of covering for a friend’s atheism by giving a long-winded answer including “as a boy he always went to church and loved singing Christmas carols” while knowing this to be untrue. In Carson’s interpretation, the whole answer constitutes evasive bullshitting while part of it is clearly a lie.

of Bender et al. of LLMs as ‘stochastic parrots’ that only repeat learned patterns of language, these models can certainly be deemed bullshitters from a linguistic perspective. They focus on form over content since the model does not ‘know’ what information is; it only ‘knows’ what information should look like. Investigating large language models as producers of bullshit texts is an important area of future research, especially once they become more common in daily life where people may use them to access factual information when the models are typically trained to reproduce surface patterns.

## 2.2 What isn’t Bullshit?

“Never tell a lie when you can bullshit your way through.” Eric Ambler’s character Arthur Abdel Simpson (cited in Frankfurt, 2005)

The term *bullshit* suffers from being a very common expletive applied to all kinds of situations. As pointed out by Bergstrom and West, not everything that makes people exclaim *That’s bullshit!* is of interest from the point of bullshit studies: “You can call bullshit on bullshit, but you can also call bullshit on lies, treachery, trickery, or injustice” (2020, p. 314). There are (at least) three different groups of phenomena that are of no interest for the current computational linguistics based analysis: non-bullshit, non-linguistic bullshit and non-pragmatic bullshit.

The first group consists of phenomena that in some way or other overlap with aspects of bullshit, lying being the prime example. Though bullshit may sometimes contain lies, Frankfurt and those who follow his line of reasoning agree that, though related, bullshitting is different from lying. Indeed, psychological research has provided evidence that at least some kinds of bullshitters are averse to lying and avoid it if possible (Littrell et al., 2021a,b). In a similar vein, simply telling the truth is also not bullshitting. On the surface, bullshit can be both true or false, but the intentions and functions of truthful assertions and bullshit statements differ greatly from a pragmatic standpoint.

Strongly related to lying, albeit created in a more organized fashion, is propaganda. Cassam rightfully points out that it seems wrong to call the propaganda in a Göbbels speech bullshit (Cassam, 2019). However, the similarities are there: Propagandists, like (Frankfurtian) bullshitters do not care whether what they are saying is true or not. Propaganda may be more effective if rooted in truth, but may just as easily appeal to people’s preconceived notions or stereotypes. What makes propaganda different from bullshit, aside from the intuition that it is somehow ‘worse,’ is its usually more concerted nature. Propaganda is often carefully crafted by multiple people (e.g. state actors) with a specific purpose in mind. It may be used to enhance a specific person’s image (e.g. in dictatorships) but it does not typically originate from that person in the spur of the moment. It therefore differs not so much in form, but rather in its origins.<sup>4</sup>

Another concept which has long been connected to bullshit is obscurantism (Cohen, 2002; Ivanković, 2016). It is certainly viable to define obscurantism as a subtype of bullshit: Typically intended to heighten the status of an author by portraying them or their work as something so complex that it is not understandable to the reader, obscurantism may constitute a kind of bald-faced bullshitting; comparable to the prosocial bullshitting in bull sessions and other settings where the language game nature of the utterances is apparent to all – or at least all who are in the know. On

4. An automated system might thus classify propaganda as bullshit, and the distinction would rather be one of taste: The term *bullshit* carries with it a certain levity that we may not want to assign to the much more negatively connotated propaganda. However, one could imagine a system that outputs all possible instances of bullshit/propaganda and human annotators could then differentiate between the two based on social and other language external factors.

the other hand, obscurantism may be interpreted as “betray[ing] an indirect move to confound while promising deep content” (Bien, 2021, p. 1498). Following this view, obscurantism would be more closely related to so-called argumentative bullshit (Gascón, 2021). Such more or less deceptive practices rely on at least some part of the audience acknowledging that epistemic conventions are lax for specific purposes: In movies or on stage, actors rely on their audience’s knowledge that what they are saying is not serious and resides outside of truthful and deceptive discourse. The same holds for humor and (in-)jokes, as well as irony, sarcasm, hyperbole and other rhetoric devices.

The second group of bullshit-related phenomena includes non-linguistic bullshit such as bullshit jobs<sup>5</sup> or bullshit visualizations. Bergstrom and West provide a number of graphics and charts that are so overloaded with fluff or simply misleading, and “attempt to be cute [making] it harder for the reader to understand the underlying data” (Bergstrom and West, 2020). Such graphics are not deceptive about their content but about the medium itself – which makes them bullshit – but they are of no particular interest to a linguist.

Lastly there are types of linguistic bullshit that will not be discussed in detail in this paper, i.e. non-pragmatic bullshit: As pointed out by Fredal, the misleading and deceptive language of George Orwell’s 1984, later called ‘doublespeak’ could be interpreted as ‘bullshit words’ in that it reinterprets words, obfuscating their conventional meaning (Fredal, 2011, p. 11). The term bullshit can also be applied to larger structures of language: “[A]rgumentative bullshit could be the production of reasons for a claim without regard to whether the reasons given really support that claim” (Gascón, 2021, p. 293), therefore using the form of an argument to deceive about it being based on sound reasoning.

## 2.3 Bullshit as Pragmatic Deception

### 2.3.1 INTUITION

In computational linguistics and NLP, much research focuses on factual deception such as lies and, especially in recent years, fake news debunking and related tasks. One standard approach consists of identifying claims and then comparing them to some sort of database for verification (see Zhou and Zafarani, 2020; Augenstein, 2021; Guo et al., 2022). For identifying bullshit statements, however, the question is not *Is this true?*, but rather *What does this actually say?*, making existing fact checking pipelines useless for bullshit detection, apart from the claim identification part.

Since the deception in bullshit is about language itself, a bullshit detection pipeline instead needs to identify the points at which pragmatic conventions or models are deceptively broken: “The normal assumptions that interlocutors make about the veracity and relevance of another’s statements (relying on Paul Grice’s maxim of Quality, for example) are misplaced when applied to the bullshitter: we think this person is having a ‘serious’ conversation when such is not the case” Fredal, p. 21. While classic pragmatic models – like the Gricean Maxims – may be useful for linguistic analysis of a single specific bullshit example, they lack a clear path towards operationalization on a larger scale. That is, they rely too much on the linguistic intuitions of annotators and world knowledge when it is highly doubtful if these can be translated into an algorithmic approach.

Computational pragmatics is a fairly new field, still finding new ways to deal with the often difficult-to-grasp aspects of communication in context. If we want to identify instances of pragmatic conventions being broken from an NLP perspective, we may first need a robust pragmatic model of cooperative dialog and build bullshit detection on top of it. As mentioned above, of the two

---

5. Which take the form of useful jobs, but are empty and meaningless (Graeber, 2013)

recent linguistic approaches to bullshit, the QUD-based approach by Stokke and Fallis seems more promising from the standpoint of operationalization. The next chapter will provide a short overview of the QUD model and how it may be adapted for bullshit analysis.

### 2.3.2 BULLSHIT AND QUDS

Over the last decades, a number of researchers have identified (implicit) underlying questions as a tool to describe the internal structure of discourse (e.g. Polanyi, 1988; von Stutterheim and Klein, 1989; Kuppevelt, 1995; Ginzburg, 1994, 1996). This paper largely follows the notion of QUDs proposed by Roberts (2012). The framework “which treats discourse as a game, with context as a scoreboard organized around the questions under discussion by the interlocutors,” (Roberts, 2012, p. 1) provides us with a way of interpreting speaker’s utterances in the dialog flow. The QUD approach revolves around the idea of a stack of so-called ‘current questions’ which are salient in the discourse. Dominated by the ‘Big Question’ (*What is the world like?*), discourse in the QUD model aims at progressively answering more and more detailed subquestions.

In this framework, cooperative speakers are beholden to answer – or more broadly to ‘deal with’ – these current questions. Current questions can be overt (e.g. *What happened to you?*) or implicit (e.g. the utterance *You look horrible!* may entail the same current question *What happened to you?*). Discourse participants may then choose to answer the question (e.g. *I fell down the stairs.*), answer a subquestion that ‘closes’ the superquestion (e.g. *Let’s just say be careful when walking down slippery stairs...*), or dismiss the question (e.g. *I don’t want to talk about it*). However, current questions cannot be simply ignored in cooperative discourse since they are highly salient and should thus be on the forefront of the participants’ minds.

Previously, the QUD framework was used to annotate discourse structure (De Kuthy et al., 2018), automatically generate potential QUDs from assertions (De Kuthy et al., 2020), as well as generating QUDs evoked by the preceding context, in essence predicting current questions and the following discourse (Westera et al., 2020). The Ginzburg et al. (2022) paper, mentioned above, also uses QUDs as a basis for analyzing the relevance of responses; building on Ginzburg (2012), the authors provide a formal analysis of both cooperative and uncooperative discourse or question-answering. The QUD framework – and not other pragmatic models that may entail deception such as Asher et al. (2017); Asher and Paul (2018) – was chosen as a basis for computational bullshit detection for two main reasons: Firstly, the definition by Stokke and Fallis is a good starting point and is already based on QUDs. Secondly, in dealing with QUDs similarly to overt questions, we may leverage the large amount of NLP research in question related fields.

From a bullshitting perspective, there are two ways in which current questions may be deceptively abused: a) providing a straight-up answer without having evidence for or caring about its truth value; and b) introducing or answering a different question portraying it as pertaining to the original one. The first option is the one discussed by Stokke and Fallis who propose the following definition of bullshitting:

- (i) *A* is bullshitting relative to a QUD *q* if and only if *A* contributes *p* as an answer to *q* and *A* is not concerned that *p* be an answer to *q* that her evidence suggests is true or that *p* be an answer to *q* that her evidence suggests is false. (Stokke and Fallis, 2017, p. 288)

Stokke and Fallis show how their definition of bullshit subsumes that of Frankfurt, but also serves to explain many of the other examples (and indeed counterexamples) which have been put



forth in the intervening years. In particular, they show that infidelity to the QUD explains cases in which bullshitters care about the truth of what they are saying, but may also lie if necessary.

Stokke and Fallis' definition in (i) does not include those (evasive) cases in which speakers stray from the original QUD and introduce (or implicitly answer) a novel QUD. Since psychological research has shown this kind of bullshitting to be distinct both cognitively and in the situations it occurs, this might be fine: There may simply be two distinct pragmatic phenomena that may call for two distinct definitions. However, the connections between the two kinds of bullshitting, as well as the shared history of research call for a definition that covers both which will be provided in the next chapter.

### 3. Operationalization

#### 3.1 Defining Bullshit

Recent psychological research indicates the importance of capturing two kinds of not caring about answering the QUD: speakers trying to appear to answer a QUD without having evidence of their answer being a truthful one (often persuasive bullshitting) or deceptively introducing or implying a new QUD (typical for evasive bullshitting). An ideal automated system should be able to capture both, as they share some important characteristics. That is, both persuasive and evasive bullshit uses the form of cooperative discourse to hide the fact that its content is 'empty' (cf. Spicer, 2020) with regard to the current discourse.

Current questions in bullshitting situations differ from those in non-bullshitting situations. For most kinds of bullshitting, QUDs are strongly connected to self-representation and image. While most communication entails subjective markers and practices of creating a self-image, bullshit seems overtly connected to it. In felicitous discourse, the Big Question should be *What is the world like?*, yet for the QUD-bullshitter, the most important discourse question may be *What am I like?*

This holds true for the bullshitting examples in section 3.1.1: Frankfurt's 4<sup>th</sup> of July Orator in example (6), Carson's Careful Test Taker in (3), Stokke and Fallis' wishful bullshitter in (5) all misuse the current (implied or overt) question under discussion to talk about themselves: *I am someone who values religion*, *I am someone who knows things – even if I can't answer this particular question* and *I am someone who wants it not to rain*, respectively. Even the more abstract cases of bullshit can be analyzed in terms of a shifted big question: Meibauer's advertising bullshit tells readers something about how a product should be perceived – not about what it factually is. Bergstrom and West's examples of overly illustrated graphs use visual means to heighten their own cleverness while (perhaps unintendedly) obscuring the actual data. For large language models, the Big Question might be *How does 'what is the world like' look like?*

These observations also hold true for linguistic bullshit, both of the persuasive and evasive kind. Evasive bullshitters may often dodge the QUD to proclaim that they are '(not) someone who does/is/believes X', but may also use the opportunity to define their self-image by appearing to answer a relevant QUD:

- (4) Q: 'Are you going to contest Roe vs. Wade?'  
 A: 'I am someone who believes in the constitution and the Supreme Court' (cf. Meibauer, 2018, p. 367).

The answer in this example is ‘empty’ with regard to the overt QUD. It answers other QUDs (*Who am I? What do I believe in?*), implying that they may be relevant subquestions that closes the original QUD. At closer inspection, it can be interpreted as *yes*, *no*, or *neither*. It only serves to build the image of the speaker by using heavily connotated phrases like ‘constitution’.

Defining bullshit as pragmatic deception thus leads to an expanded version of Stokke and Fallis’ definition:

- (ii) *A* is bullshitting relative to a QUD *q* if *A* contributes *p* as an answer to *q* and *A* is not concerned that *p* be an answer to *q* that her evidence suggests is true or that *p* be an answer to *q* that her evidence suggests is false. (Stokke and Fallis, 2017, p. 288)

*A* is also bullshitting relative to a QUD *q* if *A* introduces, or by answering implies, a novel QUD *q'*, misrepresenting it as pertaining to the original QUD *q*.<sup>6</sup>

The first part of the definition still contains the problematic (from a text analysis perspective) reliance on the speaker caring about the truth of an answer. The second part, in turn, entails misrepresentational intent. Both may result in annotators having to use subjective judgments which, as we will see later, complicates classification. However, there may exist approaches that focus on a subset of bullshit that answers questions which by definition can be neither true nor false, as in example (1). The second part may also be helpful in social media contexts, where (persuasive) bullshit often occurs unprompted, e.g. in tweets that nevertheless imply an overarching topic with associated QUDs by way of hashtags or other markers (cf. example 8). The next section shows how this definition can be operationalized to classify bullshit in future projects.

### 3.1.1 EMPIRICAL COVERAGE

Having formulated a definition that encompasses the desired types of bullshitting, the next step is the creation of a corpus. The bullshit definition in (ii) must therefore be operationalized for use by annotators. A preliminary annotation flowchart was created, consisting of a series of Yes/No questions that replicate the decision process in bullshit identification (Figure 1). An instance of Stokke and Fallis bullshit, for example, would map to YES YES YES NO (YYYN) – moving through nodes A, B, C, and D. Conversely, a simple exclamation like ‘Hi Mary!’ would map to NN (nodes A and B) and not constitute bullshit. The flowchart was used to qualitatively annotate some typical examples of bullshit.

#### (3) Careful Test Taker

A student who gives a bullshit answer to a question in an exam might be concerned with the truth of what she says. Suppose that she knows that the teacher will bend over backwards to give her partial credit if he thinks that she may have misunderstood the question, but she

6. As one anonymous reviewer pointed out, there are other, similar kinds of pragmatic deception. A common trope in spy novels has one agent telling another *I had a turbulent flight* to indicated that an adversary was on the plane with them. In my view, the difference is that such coded language (similar to the language of so-called dog whistles) relies on the intended audience knowing that there is a second, hidden QUD. The speaker does not use the surface QUD (about the tranquility of the flight) to talk about themselves, but instead using it to covertly answer the hidden QUD (about the adversary, using previously agreed upon codes). Neither Test-Taker nor teacher in (3), for example, believe that an evasive bullshitting answer pertains to some sort of secret question. Similar to in-jokes, playacting, etc, coded speech has an intended audience which is not the target of deception, while bullshitting does not.

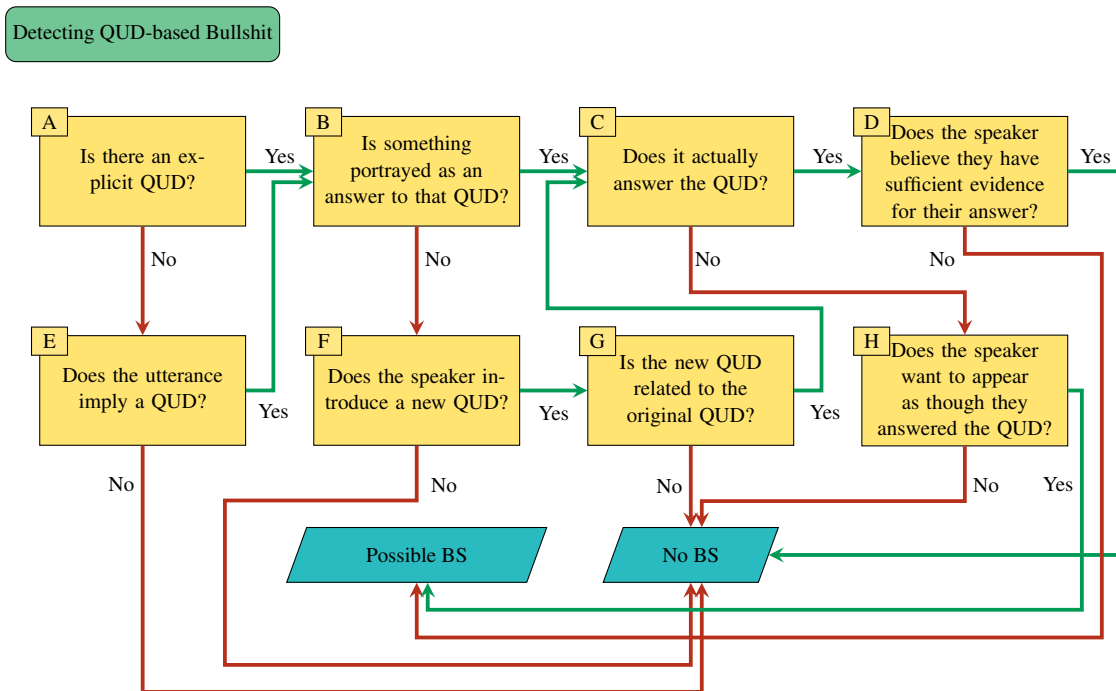


Figure 1: QUD-based Bullshit Flowchart

also knows that if the things she writes are false she will be marked down. In that case, she will be very careful to write only things that are true and accurate, although she knows that what she writes is not an answer to the question. (Carson, 2016, p. 62)

Since we have no insight into speakers' minds, many bullshit statements have several interpretations. Carson's example can be interpreted in at least two different ways.

Interpretation 1:

1. Node A: Is there an explicit QUD? → **Yes**, the teacher's question.
2. Node B: Is something portrayed as an answer to that QUD? → **Yes**, the student's answer.
3. Node C: Does it actually answer the QUD? → **No**, the student does not know the answer.
4. Node H: Does the speaker want to appear as though they answered the QUD? → **Yes**, the student provided an answer about a related topic, implying it answers the original QUD.

Path: YYNY → Possible Bullshit

Interpretation 2 (following Carson):

1. Node A: Is there an explicit QUD? → **Yes**, the teacher's question.

2. Node B: Is something portrayed as an answer to that QUD? → **No**, the student's answer is overtly meant for another question.
3. Node F: Does the speaker introduce a new QUD? → **Yes**, the seemingly misunderstood question.
4. Node G: Is the new QUD related to the original QUD? → **Yes**, the student wants to stay as close as possible to the original question to obtain partial credit.
5. Node C: Does it actually answer the QUD? → **No**, the student does not know the answer to the original question.
6. Node H: Does the speaker want to appear as though they answered the QUD? → **Yes**, the student wants to appear as though they answered the question they misunderstood, even though it differs from the actual question asked by the teacher.

Path: YNYYNY → Possible Bullshit

(5) **Wishful bullshitter**

Jack and Julia are going to Chicago. They have tickets to a Cubs game, and being a big Cubs fan, Julia hopes the game will not be rained out. A few days before their departure, they are talking about their trip. Jack says, 'I'm really looking forward to that Cubs game. I hope it won't rain.' Julia replies with a confident air, 'This time of year, it's always dry in Chicago.' But she has no evidence about the weather in Chicago, and she has no idea whether it's likely to rain or not. (Stokke and Fallis, 2017, p. 7)

1. Node A: Is there an explicit QUD? → **No**
2. Node E: Does the utterance imply a QUD? → **Yes**, whether or not it will rain in Chicago.
3. Node B: Is something portrayed as an answer to that QUD? → **Yes**, Julia's assertion that it is always dry this time of year.
4. Node C: Does it actually answer the QUD? → **Yes**.
5. Node D: Does the speaker believe they have sufficient evidence for their answer? → **No**, Julia only wishes it to be so.

Path: NYYYN → Possible Bullshit

(6) **4<sup>th</sup> of July Orator**

Consider a Fourth of July orator, who goes on bombastically about 'our great and blessed country, whose Founding-Fathers under divine guidance created a new beginning for mankind.' (Frankfurt, 2005, p. 4)

1. Node A: Is there an explicit QUD? → **No**
2. Node E: Does the utterance imply a QUD? → **Yes**, whether or not America was founded under divine guidance.
3. Node B: Is something portrayed as an answer to that QUD? → **Yes**, the orator's assertion both implies the QUD and affirms it.
4. Node C: Does it actually answer the QUD? → **Yes**.

5. Node D: Does the speaker believe they have sufficient evidence for their answer? → **No**, the speaker has no insight into the divine (in Frankfurt's interpretation).

Path: NYYYN → Possible Bullshit

(7) **Trump & Putin**

Reporter:

Do you accept that part of the finding? And will you undo what President Obama did to punish the Russians for this or will you keep it in place?

Donald Trump:

Well, if – if Putin likes Donald Trump, I consider that an asset, not a liability, because we have a horrible relationship with Russia. Russia can help us fight ISIS, which, by the way, is, number one, tricky. I mean if you look, this administration created ISIS by leaving at the wrong time. The void was created, ISIS was formed. If Putin likes Donald Trump, guess what, folks? That's called an asset, not a liability. Now, I don't know that I'm gonna get along with Vladimir Putin. I hope I do. But there's a good chance I won't. And if I don't, do you honestly believe that Hillary would be tougher on Putin than me? Does anybody in this room really believe that? Give me a break. (Trump, 2017)

1. Node A: Is there an explicit QUD? → **Yes**, the reporter asked two overt questions.
2. Node B: Is something portrayed as an answer to that QUD? → **No**, the overt questions are not directly addressed.
3. Node F: Does the speaker introduce a new QUD? → **Yes**, Donald Trump introduces several new QUDs, e.g. *Is it an asset that Putin likes Trump?* and *Would Hillary be tougher on Putin?*.
4. Node G: Is the new QUD related to the original QUD? → **Yes**, the new QUDs all have to do with Trump's relationship with and punishment of Putin.
5. Node C: Does it actually answer the QUD? → **No**, e.g. the question aimed at particular sanctions by the previous president is not addressed.
6. Node H: Does the speaker want to appear as though they answered the QUD? → **Yes**, by implying that he will be tough(er than Hillary), Trump wants to connect to the original QUD about "punish[ing] the Russians".

Path: YNYYNY → Possible Bullshit

(8) **Social Media**

Corona Virus is Temporary. House music is forever 🙌 (MATRODA [@matrodamusic], 2020)

1. Node A: Is there an explicit QUD? → **No**
2. Node E: Does the utterance imply a QUD? → **Yes**, the parallelism implies two QUDs about how long both the corona virus and house music will last.
3. Node B: Is something portrayed as an answer to that QUD? → **Yes**, the tweet author's assertion both implies the QUDs and affirms them.

Name 3 important historical events that happened on August 12th

1. In 1519, the first known European landing on the mainland of the Americas occurred when Spanish conquistador Hernán Cortés set foot on what is now Mexico. This event is considered one of the most significant in world history, as it marked the beginning of the widespread European exploration and colonization of the New World.
2. In 1793, the French Revolution's Reign of Terror began, during which thousands of people were arrested, imprisoned, and executed for perceived opposition to the revolution.
3. In 1953, the Korean War armistice was signed, ending three years of fighting between North and South Korea. The armistice, which was signed by the United States, North Korea, and China, established a demilitarized zone between the two countries and a cease-fire, but no peace treaty was signed, and the Korean War is technically still ongoing.

Figure 2: Example of LLM Bullshit

4. Node C: Does it actually answer the QUD? → **Yes**.
5. Node D: Does the speaker believe they have sufficient evidence for their answer? → **No**, the speaker has no insight into the longevity of a novel virus and a style of music.

Path: NYYYN → Possible Bullshit

#### (9) ChatGPT

The example of ChatGPT bullshit in Figure 2 was generated on January 9<sup>th</sup> 2023 with the December 15<sup>th</sup> release of ChatGPT (<https://chat.openai.com/>).<sup>7</sup>

1. Node A: Is there an explicit QUD? → **Yes**, The prompt written by the author.
2. Node B: Is something portrayed as an answer to that QUD? → **Yes**, the chatbot provides three examples as asked.
3. Node C: Does it actually answer the QUD? → **Yes**, it directly – but incorrectly – answers the question.
4. Node D: Does the speaker believe they have sufficient evidence for their answer? → **No**, the system by its design cannot believe anything and simply outputs something that fits the shape of a perfect answer (following the reasoning of Bender et al., 2021).

Path: YYYN → Possible Bullshit

7. ChatGPT bullshit, arguably, could be seen as prompted and persuasive. The system has no mechanism to try to evade a question, rather it does not – because it cannot – care about the content as in persuasive bullshit.

These examples show that manual, in-depth analysis of common bullshitting examples is possible and some of the nodes may even be automated and handled with NLP methods. Nodes A and B and F fall squarely into the space of question and answer detection. Node E will mostly lead to YES for assertions, since most utterances at least imply a topic (and assertion detection is reasonably simple), which can then be identified via topic classification methods which also applies to Node G. Similarly, answers and newly introduced QUDs may be identified with computational methods. Node C, *Does it actually answer the QUD?* is a classic NLP problem for question answering, but it is still far from solved. Nodes D and H, though, are highly problematic from an NLP standpoint. The next chapter will discuss some limitations and caveats of this detailed annotation process for bullshit examples.

### 3.1.2 LIMITATIONS AND CAVEATS

While the QUD approach to bullshit may at first glance seem like a promising avenue of research, there are still some issues: The first is that QUD annotation is a fairly recent practice, though various guidelines for QUD annotation have been provided (see De Kuthy et al., 2018; Riester et al., 2018; Riester, 2019). Second, even the simplified bullshit annotation process presented in the previous chapter leaves much to the interpretation of the annotators. Discourse often depends on context and audience so propositions and current questions may be interpreted in any number of ways. As mentioned with example (1) and also noted by Fredal: “audiences will differ in their response to a speaker’s statements and motives, some seeing truth and honesty where others see various degrees of bias, deception, and misinformation” (Fredal, 2011, p. 20). It is doubtful whether a high inter-annotator agreement can be reached under such conditions, so the task may be one that has to rely on differing, subjective labels.

Third, and from a computational perspective maybe most concerning, are the kinds of nodes in the QUD annotation flowchart that rely on language-external knowledge, especially *Does the speaker believe they have sufficient evidence for their answer?* (Node D) and to a certain degree *Does the speaker want to appear as though they answered the QUD?* (Node H). While the latter may be supported by some evidence on the linguistic level (e.g. by the use of specific words or grammatical markers of deception), the former requires insight into the speakers’ minds. Though there might be linguistic markers that hint at a speaker’s degree of evidence, in the majority of cases there should be no way of discerning it by looking at the text. Especially when taking into account Meibauer’s notion of ‘too much certainty,’ the whole point is that the linguistic markers (of certainty) do not match the speaker’s internal state of mind or depth of knowledge.

We may hope to find solutions in annotating large amounts of texts containing bullshit, using linguistically and epistemologically trained annotators. NLP methods such as large language models or ‘foundation models’ (Bommasani et al., 2021) may then be employed to find latent linguistic markers that serve as indicators of the authors’ internal state of mind – if such markers exist. However, it is doubtful that such an approach will work, especially with the kinds of short texts in social media or impromptu interview answers.<sup>8</sup> A more fruitful endeavor, for now, may be trying to identify breaking points in pragmatic models in other, more straight-forward ways. One such practice of pragmatic deception – in the form of *shifting* – will be discussed in the next chapter.

---

8. In addition to a range of other concerns regarding the use of LLMs, as noted by Bender et al. (2021) and others.

### 3.2 Investigating Shifting as a Practical Way of Finding Pragmatic Deception

Approaches based on holistic pragmatic models may well remain in the area of theoretical linguistics, so for the foreseeable future, NLP research must focus on detecting specific bullshit subtypes. One example is *shifting*, a type of evasive answers introduced in an journalistic analysis of Danish Prime Minister Lars Løkke Rasmussen, whose was frequently criticized for evasive behavior in interviews (Gabrielsen et al., 2020). Rather than employing techniques shown in Bull and Mayer (1993) like verbally attacking the interviewer or questioning the question, Rasmussen evades by shifting not the topic (as in Greatbatch, 1986), but something else. Though the authors do not use the term, shifting may be interpreted as evasive bullshitting, since the speaker “stays within the topical agenda set by the question [...] but refocuses to a more favorable aspect within that topic” (Gabrielsen et al., 2020, p. 1355). The authors do not use the QUD framework, but their concept of shifting fits well within the QUD-based definition of bullshitting in (ii) in section 3.1: “When a shift is successfully executed, it will appear as if the interviewee answers the question when in fact the interviewee has steered clear of the critical aspect of the question thereby leaving the original question unanswered” (Gabrielsen et al., 2020, p. 1356). Fact-checking does not reveal shifting, because it does not involve giving false information or changing the topic. Instead, the interviewee changes the focus of their answer, interpreting what is considered relevant within that topic (Gabrielsen et al., 2020, p. 1356f).

The authors identify three different types of shifts: shift of time, agent, and level, while acknowledging that there may be many more. Shifts of time happen when the question concerns a specific time frame and the answer is shifted to a different one. When asked about specific economic policy in the past, e.g., the Danish Prime Minister instead talked about “the present economic situation as well as future expectations for the economy” (Gabrielsen et al., 2020, p. 1362). Shifts of agent can involve broadening the agent, i.e. when Rasmussen referenced his party, government or the country in answers to questions that are directed at his own opinions or actions (Gabrielsen et al., 2020, p. 1363). Agents can also be narrowed, when replying to questions like *What is the party’s opinion on supporting families?* with the agent shifted to a narrower scope like *As a parent, I think that...* (cf. Gabrielsen et al., 2020, p. 1363). Shifts of level lead to answers that are either more abstract or more concrete than asked for. Most commonly, this occurred when the Prime minister shifted a question about concrete politics to underlying, but more abstract ideological motives (Gabrielsen et al., 2020, p. 1365).

These types of shifts – and others yet to be identified – are clear examples of pragmatic, evasive bullshitting. Instead of rejecting the question (or QUD) outright, they reject “the journalist’s underlying intention with the question” (Gabrielsen et al., 2020, p. 1367). That is, they imply novel QUDs misrepresenting them as pertaining to the original QUD as described in definition (ii) in Section 3.1. These shifts are difficult to detect in the moment and in the dataset, no journalist called the Prime Minister out on this behavior.

Nevertheless, the authors note a number of linguistic indicators of shifting, which is especially interesting from an NLP perspective, since evidence of this deceptive practice may be found in the text, without insight into the speaker’s mind. Shifts of time may lead to mismatched temporal adjectives and grammatical tense in questions and corresponding answers, shifts of agents to mismatching pronouns or verb forms. Shifts of level may manifest on the linguistic surface in the form of more concrete or abstract adjectives, verbs, nouns, etc. If they prove robust, such verbal



indicators could be leveraged for automatic classification by use of language models or other forms of machine learning, which makes investigating these markers vital for shifting annotation.

### 3.2.1 ANNOTATING SHIFTS

To investigate whether approaching (a subset of) bullshit from an NLP perspective becomes feasible when focusing on shifting, we created a small test corpus. The source for the corpus was the German Parliament’s question hour (‘Befragung der Bundesregierung’)<sup>9</sup>, which is a prototypical bullshitting scenario, as it comprises situations in which persons are strongly expected to answer but may not want to do so for any number of reasons. 100 sample question answer pairs were selected from random parliamentary sessions of the past eight years.

To ascertain whether the notion of evasive bullshitting in general and shifting in particular is intuitive, annotations were carried out by 26 untrained graduate students of German language and literature at the Ruhr-University-Bochum. The students were given the task of carefully reading the Gabrielsen et al. paper, then randomly assigned 15 question-answer-pairs, and asked to mark whether a given pair contains any of the three shifts (time, agent, level), or any other shift. For each Q&A pair, between three and five students handed in annotations, resulting in a total of 89 annotated pairs.

For some examples, the annotation decisions were fairly straightforward. In example (10), it is immediately apparent that the speaker does not attempt to evade the question. Instead, they openly admit that they cannot answer the question and refer it to someone else.<sup>10</sup> In such a case, we need not even look at the question, since such an answer can neither be shifting nor any other kind of QUD-based bullshit. In example (11), on the other hand, we need to look at the question to see that the answer contains a shift of level in talking about abstract concepts (i.e. *We appreciate the United States as a country of democracy and that is the basis of international diplomacy*) when the questions were rather specific (*Do you believe today in a joint and, above all, meaningful closing statement?*).<sup>11</sup>

- (10) Da ich mich dort ehrlicherweise nicht mit den Details auskenne, würde ich den zuständigen Kollegen – wahrscheinlich Ressort Umwelt oder Wirtschaft – bitten, entsprechend zu antworten.

Since I am honestly not familiar with the details there, I would ask the responsible colleague – probably Department of Environment or Economy – to answer accordingly. (Richter et al., 2020, ID 1011958)

- (11) **Frage**  
Frau Bundeskanzlerin, welchen Sinn hat ein Gipfeltreffen im Rahmen der G 7, wenn Sie und nahezu alle Mitglieder Ihrer Bundesregierung den amerikanischen Präsidenten fortgesetzt und auf allen öffentlichen Kanälen diskreditieren? Glauben Sie heute angesichts der vor sich hergetragenen Vorurteile an eine gemeinsame und vor allem sinnvolle Abschlusserklärung?

9. Data accessed via the Open Discourse Corpus (Richter et al., 2020).

10. Of course, the answer could still be a lie if the speaker is in fact familiar with the details, but it is not a case of bullshitting.

11. The analysis is complicated by the fact that the original speaker asks leading questions that serve other pragmatic goals and which almost call for a shift in order to answer without damaging one’s own position. Whether there is such a thing as a ‘bullshit question’ remains to be seen (see also Meibauer, 2016, p. 84). For further insight into types of questions in political settings, see Zhang et al. (2017).

Und sind unter diesen Umständen die wieder einmal massiven und wohl auch kostenintensiven Sicherheitsmaßnahmen auf diesem Gipfel gerechtfertigt? (Richter et al., 2020, ID 1007478)

**Antwort**

Sie haben eine hohe Bandbreite an Fragen, die Sie aus Ihrer Fraktion heraus stellen, was das Verhältnis zu den einzelnen Ländern anbelangt. Wir schätzen die Vereinigten Staaten als ein Land der Demokratie, aber wir sind trotzdem der Meinung, dass da, wo Meinungsverschiedenheiten bestehen, diese auch benannt werden müssen. Aber gerade in solchen Zeiten ist es eben einfach auch wichtig, immer wieder den Gesprächsfaden zu suchen und Überzeugungsarbeit zu leisten – darauf beruht internationale Diplomatie –, und das tun wir in alle Richtungen. (Richter et al., 2020, ID 1007479)

**Question**

Madam Chancellor, what is the point of a G-7 summit meeting if you and almost all members of your federal government continue to discredit the American president on all public channels? Do you believe today in a joint and, above all, meaningful closing statement in view of the prejudices you have displayed? And under these circumstances, are the once again massive and probably costly security measures at this summit justified?

**Answer**

You have a large range of questions that you ask out of your parliamentary group in terms of the relationship with the individual countries. We appreciate the United States as a country of democracy, but we are nevertheless of the opinion that where there are differences of opinion, these must also be named. But it is precisely in times like these that it is important to keep trying to find the thread of dialogue and to persuade others – that is the basis of international diplomacy – and we do that in all directions.

Since no clear consensus was reached by the students for many of the QA pairs, all were annotated a second time by two trained annotators, including the author, discussing each pair in detail. This was done to investigate whether an in-depth analysis of each sample would lead to similar results to the more intuitive approach of the untrained students. Since Gabrielsen et al. only provided a very small number of examples, this annotation also served to investigate whether the concept of shifting is reproducible in expert annotation from a different field (linguistics as opposed to journalism).

When distinguishing shifting from non-shifting behavior, the annotators often went back to the QUD-based definition of (evasive) bullshitting as a basis of discussions. For example, in some cases the question (under discussion) was explicitly marked as unanswerable or not worth answering by the government member as in example (10). For some shifts, it was helpful to compare the overt QUD in the question to the implicit one reconstructed from the answers, to see which elements are shifted. More difficult pairs also benefited from QUD-based analysis along the lines of the flowchart in section 3.1.1, which provided clarification for otherwise more subjective judgments.

### 3.2.2 RESULTS, LIMITATIONS AND CAVEATS

Students identified at least one kind of shift in most of the examples. Using MACE (Hovy et al., 2013) as an analysis tool<sup>12</sup>, we found only 20 question-answer pairs without any shift, i.e. some

12. MACE aggregates annotations to recover the most likely answer, calculates which annotators are trustworthy and evaluates item and task difficulty. More information can be found at <https://github.com/dirkhovy/MACE>.

students might have annotated a shift in them, but the weighted consensus did not. Most annotated shifts were of the ‘shift of agent’ category, with ‘shift of level’ being a close second. ‘Shift of time’ was annotated to a lesser degree and while some students marked some ‘other’ shifts, these were never enough to lead to a consensus.<sup>13</sup> Still, in 69 Q&A pairs the consensus identified at least one of the three shifts, confirming the assumption that the question hour is indeed a genre in which evasive bullshitting or shifting occurs.

However, the pilot study indicates that a short primer on shifting is not enough to enable untrained annotators to come to a significant consensus on the various shift types. There is a reasonable annotator agreement that at least some kind of shift occurs in the majority of samples with a mean annotator competence of 0.68 when looking at whether at least one type was noted. Unfortunately, there is significantly less overlap for the types of shifts, see Table 1. Average annotator competence as provided by MACE was 0.53 for shift of time, 0.52 for shift of agent and 0.41 for shift of level. Average agreement for the ‘other’ category of shifts was 0.85 since most students did not annotate any of these shifts at all. Krippendorff alpha lay between 0 and 0.27 for all categories, indicating that annotators did not intuitively come to the same conclusions.

On the one hand, the low agreement could indicate that the task is not intuitive and that more training is needed. However, it stands to reason that bullshitting is part of a group of phenomena that are inherently subjective. We might therefore be unable to calculate a gold standard since different people will view different statements as either bullshit or not, depending on extralinguistic factors such as their view of the speaker or their general attitude towards epistemic norms as outlined in Plank (2022).<sup>14</sup> Bullshit classification may therefore benefit from taking into account different labels of all annotators since subjective information may get lost when using binary labelling.

Table 1: Average MACE Annotator Competence for Shift Annotation

| Shift of Time | Shift of Agent | Shift of Level | Other Shift |
|---------------|----------------|----------------|-------------|
| 0.53          | 0.52           | 0.41           | 0.85        |

As mentioned in Section 3.2, shifting annotation favors NLP classification of bullshit phenomena mainly because of identifiable linguistic surface markers. However, during the in-depth annotation of all question answer pairs, challenges became apparent when relying on specific words. For example, shifts of agents were surmised to be apparent in the usage of pronouns, but when a question is addressed at *you* (in both English and formal German), the answer can contain either *I* or *we*. It thus depends on annotator discretion, whether the question aimed at singular you or plural you. Similarly, there can be no indication of a shift of agent if the question does not contain a presumed agent. That is, questions like *Will the law be beneficial for...?* can be answered with pronouns or nouns for any number of agents.

We also encountered challenges when annotating shifts of level. Since the Bundestag data contains rather long answers to multiple questions at once, there is much more space to express both concrete and abstract concepts. Meaning, while the question may be concrete, the answer can contain either abstract, concrete, or mixed concepts. Concrete examples were used in combination

13. This may well be an issue of motivation for doing extra work, since the students that annotated the most ‘other’ shifts were also the ones that provided the most in-depth comments for their reasoning.

14. I am thankful for the anonymous reviewer who pointed out that some linguistic phenomena cannot and probably should not be reduced to a singular gold label.

with abstract concepts to fully answer the question and give additional context. Annotation was therefore much more complicated than just looking at whether or not abstract or concrete concepts were expressed linguistically in both question and answer or not.

It therefore remains doubtful whether textual markers are enough to classify shifts and whether an automated system will come to the right conclusions simply on the basis of the utterances themselves. A simple word-based or otherwise rule-based classification system will not be able to ignore the extraneous and misleading tokens and deep learning and architectures might not be able to pick up on nuances that require intense discussion among human annotators simply from looking at the text. Some preliminary findings are provided in the next chapter.

## 4. NLP Approaches, Conclusion & Outlook

### 4.1 NLP Methods for Bullshit Analysis

A thorough investigation of NLP methods for bullshit detection is beyond the the scope of this paper and will be approached in future research. Instead, two different possible pipelines for bullshit detection will be outlined in this chapter: one for persuasive and one for evasive bullshit. Building on the QUD-based definition of bullshitting, we can make use of established NLP methods that deal with questions in a broader sense. Yet we still face challenges in decisions human annotators make based on context, world knowledge and intuition about the producers of bullshit. The following pipelines can thus only serve as a starting point of classification for some types of bullshitting behaviors. Others remain dependant on human intervention, possibly as a final step following an automated pre-selection of bullshit candidates using NLP methods.

#### 4.1.1 NLP-BASED DETECTION OF PERSUASIVE BULLSHIT

Since persuasive bullshitting typically occurs unprompted, i.e. not following an overt question, the first step in the pipeline is claim detection. This subtask is vital for fake news detection and related fields, so we can build on a vast amount of research (e.g. Levy et al., 2014; Hassan et al., 2015; Lippi and Torroni, 2015; Barrón-Cedeño et al., 2020; Konstantinovskiy et al., 2021). Next, the system must infer the implicit questions, building on QUD-specific research by De Kuthy et al. (2020), but also on a large amount of literature on question generation for a variety of purposes (Heilman and Smith, 2010; Du et al., 2017; Duan et al., 2017; Zhou et al., 2018; Kurdi et al., 2020).

The third step in persuasive bullshit detection determines whether the generated questions are ‘unanswerable’ as in example (1) about a person’s degree of racism. A rule-based approach might suffice for many cases, based on identifying vague or unquantifiable concepts about which one cannot make assertions with sufficient certainty. Another approach could utilize neural networks either by training on corpora of unanswerable questions prepared by annotators, or by using the intuition of a large language model. While systems like ChatGPT may themselves be bullshitters – at least when it comes to content – they can show surprising accuracy on what is quantifiable: When asked, ChatGPT tells us that *weight* is quantifiable while *racism* is not. Whether we can actually rely on a system that only ‘knows’ about form to give sufficiently accurate accounts of answerability requires in-depth research in the future.

Finally, there are those cases of persuasive bullshitting in which the concept is not unquantifiable. Here, NLP methods reach their limits and human annotators will be needed to make the final decision on whether the utterance can be called bullshit. There exist NLP methods to figure out

whether the assertion fits the reconstructed implicit question, e.g. by calculating the question-answer congruence (Faruqui and Das, 2018; Yu and Jiang, 2021). In this, we must be extremely cautious to avoid circular reasoning: If the question is generated based on a piece of text, the piece of text must in some way answer it. Whether it ‘really’ fits may require linguistic and world-knowledge intuition on the part of the annotators.

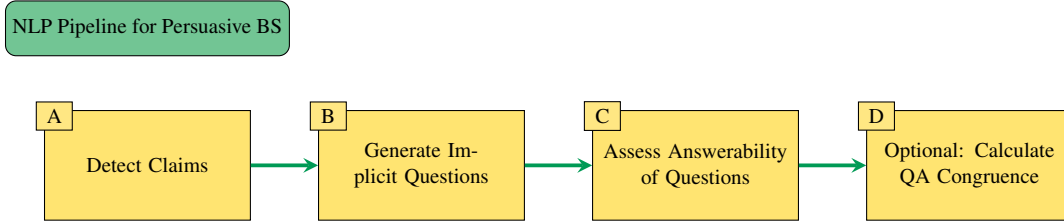


Figure 3: NLP-based Detection of Persuasive Bullshit

#### 4.1.2 NLP-BASED DETECTION OF EVASIVE BULLSHIT

Since evasive bullshit typically occurs in prompted situations, the first step is not claim detection but rather identifying the answer(s) to overt questions. Next, topic classification methods (e.g. Wang and Manning, 2012; Quercia et al., 2012; Pappagari et al., 2019) may be used on both question and answer to check if their topics match. If they do not, we can rule out evasive bullshitting, where only aspects of the question are shifted, but not the whole topic as in Greatbatch (1986). Differing topics could also stem from other types of uncooperative behavior, as shown in Bull and Mayer (1993), or simple misunderstandings.

Implicit QUDs can then be generated from the question, analogously to the detection of persuasive bullshit above. A number of NLP methods, e.g. adapted from various text similarity measures, can then be employed to compare the reconstructed QUDs to the overt question to see whether and where they overlap (Croft et al., 2013; Prasetya et al., 2018; Prakoso et al., 2021). Evasive bullshitting by way of introducing novel QUDs (and pretending they answer the original one) occurs when the topic of question and answer matches, but the implicit QUDs differ strongly from the overt ones. When the implicit QUD is very similar to the overt question, yet changed in agent, time or level of abstraction, it may constitute a case of shifting as described in Section 3.2.

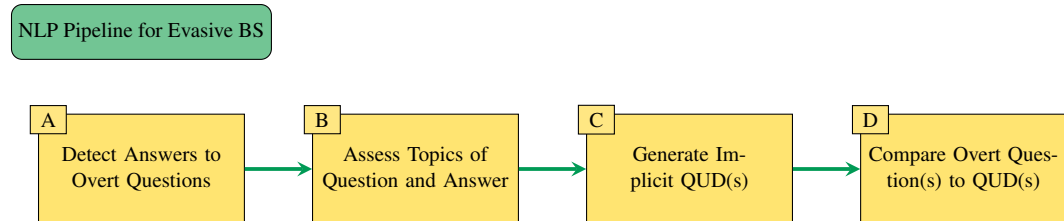


Figure 4: NLP-based Detection of Evasive Bullshit

## 4.2 Conclusion

If we want to deal with subjective and amorphous subjects such as bullshit computationally, we need to find robust definitions. We need to approach smaller, more manageable subtypes of bullshit to create corpora and build automated systems. However, finding such narrow definitions remains a challenge. As shown above, even picking only three specific sub-concepts, the three shifts, of only one type of bullshitting (evasive bullshit) does not facilitate straightforward annotation. The cooperative annotation with trained annotators provided better results, but it is also extremely time-consuming.

Pragmatic theories of bullshitting serve as starting points from which we can pick parts to be tackled with NLP methods. Annotation guidelines that cover the whole spectrum of bullshit suffer from their reliance on language-external factors. These environmental aspects of specific discourse situations and speaker-internal mental processes are exceedingly hard to grasp solely on the basis of text. According to the prevailing theories on the language of deception and lying, at least some of these internal workings are expressed linguistically (see e.g. Meibauer, 2019). It remains unclear to what degree this notion applies to the language of bullshit.

The in-depth shifting annotation with two trained annotators provided insights into this specific subtype of evasive language and that whether or not a shift can occur strongly depends on the type of question. Some questions permitted fewer shifts, e.g. when politicians are simply asked about their opinion on some topic, their answer may refer to past, present or future, so a shift of time is improbable. Shifts of time are still possible, e.g. by answering *In the past I was of the opinion that...*, but this would be a very unusual, highly marked answer in which the shift became so obvious as to be useless to the shifter.

Research strongly suggests that bullshit exists outside of the conventional truth/lie spectrum. Evasive bullshitting in the form of shifting can be found in politicians' answers; other forms of evasive bullshitting can be found anywhere from exam situations to post-game interviews – bullshit “is unavoidable whenever circumstances require someone to talk without knowing what he is talking about” (Frankfurt, 2005, p. 15). Disciplines ranging from journalism to psychology have provided evidence which strongly points toward the existence of some sort of apathy towards ones answers or some sort of emptiness with regard to a QUD, which speakers may deceive about. They do not hide the truth value of a factual concept, but rather hide their attitude towards the conventions of discourse itself. Instead of a simple disregard of truth, there is a disregard for the structures and norms of human communication which places linguistic bullshit strongly in the space of pragmatics.

Choosing Questions under Discussion as a starting point – mainly because of existing NLP research connected to questions – this paper has focused on three research questions. RQ1, *What is a useful bullshit definition for a computational linguistics approach?*, was answered in Section 3.1 by providing an extended, QUD-based definition that builds on Stokke and Fallis (2017) and covers both persuasive and evasive bullshitting. In Sections 3.1.1 and 3.2, two different approaches were provided to answer RQ2, *How can we annotate and analyze bullshit?*. Finally, RQ3, *What are promising NLP approaches to bullshit classification?*, was tentatively answered in Section 4.1. Two NLP pipelines for persuasive and evasive bullshitting detection were outlined to serve as a basis for future work.

### 4.3 Outlook

Further work is needed in identifying evasive bullshitting in prompted situations, as well as finding novel ways to tackle persuasive, unprompted bullshit. Advances in computational QUD analysis facilitate the identification of underlying questions for assertions of any kind, including bullshit statements. The advantage of relying on QUDs as an established pragmatic model of communication is the model’s relationship with (overt) questions. Linguistically, bullshit detection can also be based on other frameworks, which may be more apt to model uncooperative dialog (such as Asher et al., 2017; Asher and Paul, 2018). However, if it is possible to identify underlying questions and their answers in bullshitting texts, we may benefit from such NLP tasks as question answering or answer quality estimation. The QUD framework therefore makes the novel task of computational bullshit detection somewhat more approachable. Future work will focus on creating corpora of different types of bullshitting behavior and using the outlined NLP pipelines for bullshit detection.

In the coming years, bullshit detection will become vital from a practical perspective: Question answering system found in popular speech assistants like Apple’s Siri or Amazon’s Alexa, for example, may deem an answer as not sufficient to the speaker’s intent (as opposed to being plain wrong). Especially with the recent explosion of interest in LLM-based text generation, the concept of a bullshit answer or statement becomes increasingly important. With ethical discussions about the biases inherent in the data (see e.g. Raji et al., 2020; Bender et al., 2021; Gebru et al., 2021), philosophical questions of what these LLMs can ‘know’ when learning only on text, and practical considerations of reliability, the future of text generation will be treacherous if these systems turn out to be plain bullshitters.

## 5. Acknowledgements

I would like to thank the student annotators of shifting examples, including Julius Kirschner who helped with the in-depth annotation. I would also like to thank my anonymous reviewers and the board of Dialogue & Discourse for their invaluable feedback. Finally, I give thanks to my supervisor Tatjana Scheffler for her input and support during many discussions of bullshitting phenomena.

## References

- Nicholas Asher and Soumya Paul. Strategic Conversations Under Imperfect Information: Epistemic Message Exchange Games. *Journal of Logic, Language and Information*, 27(4):343–385, December 2018. ISSN 0925-8531, 1572-9583. doi: 10.1007/s10849-018-9271-9. URL <http://link.springer.com/10.1007/s10849-018-9271-9>.
- Nicholas Asher, Soumya Paul, and Antoine Venant. Message Exchange Games in Strategic Contexts. *Journal of Philosophical Logic*, 46(4):355–404, August 2017. ISSN 0022-3611, 1573-0433. doi: 10.1007/s10992-016-9402-1. URL <http://link.springer.com/10.1007/s10992-016-9402-1>.
- Isabelle Augenstein. *Towards Explainable Fact Checking*. Doctor Scientiarum (Dr. Scient.), University of Copenhagen, December 2021. URL <http://arxiv.org/abs/2108.10274>.
- Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. Proppy: A System to Unmask Propaganda in Online News. *Proceedings of the AAAI Conference on Artificial*

- Intelligence*, 33:9847–9848, July 2019a. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33019847. URL <https://www.aaai.org/ojs/index.php/AAAI/article/view/5061>.
- Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, September 2019b. ISSN 03064573. doi: 10.1016/j.ipm.2019.03.005. URL <https://dx.doi.org/10.1016/j.ipm.2019.03.005>.
- Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névél, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 215–236, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58219-7. doi: 10.1007/978-3-030-58219-7\_17. URL [https://dx.doi.org/10.1007/978-3-030-58219-7\\_17](https://dx.doi.org/10.1007/978-3-030-58219-7_17).
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL <https://dl.acm.org/doi/10.1145/3442188.3445922>.
- Carl T. Bergstrom and Jevin D. West. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. Random House, New York, first edition edition, 2020. ISBN 978-0-525-50918-9 978-0-593-22976-7. URL <https://www.callingbullshit.org/>.
- Josh Bernoff. ChatGPT is a bullshitter, December 2022. URL <https://withoutbullshit.com/blog/chatgpt-is-a-bullshitter>.
- Eric Nenkia Bien. How Obscurantism Differs from Bullshit: A Proposal. *Theoria*, 87(6):1497–1526, December 2021. ISSN 0040-5825, 1755-2567. doi: 10.1111/theo.12354. URL <https://onlinelibrary.wiley.com/doi/10.1111/theo.12354>.
- Rishi Bommasani, Drew A. Hudson, E. Adeli, R. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, E. Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, S. Ermon, J. Etchemendy, Kawin Ethayarajh, L. Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, S. Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Dan Jurafsky, Saahil Jain, Pratyusha Kalluri, Siddharth Karamcheti, G. Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, M. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, J. Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, A. Narayan, D. Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles,



- H. Nilforoshan, J. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, J. Park, C. Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R’e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, K. Srinivasan, Alex Tamkin, Rohan Taori, A. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, M. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, August 2021. URL <https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models>.
- Peter Bull and Kate Mayer. How Not to Answer Questions in Political Interviews. *Political Psychology*, 14:651, December 1993. doi: 10.2307/3791379.
- Thomas L. Carson. Frankfurt and Cohen on bullshit, bullshitting, deception, lying, and concern with the truth of what one says. *Pragmatics & Cognition*, 23(1):53–67, September 2016. ISSN 0929-0907, 1569-9943. doi: 10.1075/pc.23.1.03car. URL <https://dx.doi.org/10.1075/pc.23.1.03car>.
- Quassim Cassam. The Bullshit Industry, November 2019. URL <https://www.quassimcassam.com/talks>.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL <https://aclanthology.org/D18-1241>.
- Gerald Allen Cohen. Deeper into Bullshit. In *Contours of Agency: Essays on Themes from Harry Frankfurt*, pages 321–339. MIT Press, 2002. ISBN 978-0-262-52813-9. URL <https://mitpress.mit.edu/books/contours-agency>.
- David Croft, Simon Coupland, Jethro Shell, and Stephen Brown. A fast and efficient semantic short text similarity metric. In *2013 13th UK Workshop on Computational Intelligence (UKCI)*, pages 221–227, September 2013. doi: 10.1109/UKCI.2013.6651309.
- Giovanni Da San Martino, Chris Brew, Giovanni Luca Ciampaglia, Anna Feldman, Chris Leberknight, and Preslav Nakov. Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Barcelona, Spain (Online), December 2020a. ISBN 978-1-952148-36-1ii. URL <https://aclanthology.org/volumes/2020.nlp4if-1/>.
- Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A Survey on Computational Propaganda Detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4826–4832, Yokohama, Japan, July 2020b. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/672. URL <https://dx.doi.org/10.24963/ijcai.2020/672>.

- Kordula De Kuthy, Nils Reiter, and Arndt Riester. QUD-Based Annotation of Discourse Structure and Information Structure: Tool and Evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, January 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1304>.
- Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. Towards automatically generating Questions under Discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.509. URL <https://www.aclweb.org/anthology/2020.coling-main.509>.
- Xinya Du, Junru Shao, and Claire Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1123. URL <https://aclanthology.org/P17-1123>.
- Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question Generation for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1090. URL <https://aclanthology.org/D17-1090>.
- Manaal Faruqui and Dipanjan Das. Identifying Well-formed Natural Language Questions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 798–803, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1091. URL <https://aclanthology.org/D18-1091>.
- Harry G. Frankfurt. *On Bullshit*. Princeton University Press, Princeton, NJ, 2005. ISBN 978-0-691-12294-6. URL <https://www.jstor.org/stable/j.ctt7t4wr>.
- James Fredal. Rhetoric and Bullshit. *College English*, 73(3):243–259, 2011. ISSN 0010-0994. URL <https://www.jstor.org/stable/25790474>.
- Jonas Gabrielsen, Heidi Jønch-Clausen, and Christina Pontoppidan. Answering without answering: Shifting as an evasive rhetorical strategy. *Journalism*, 21(9):1355–1370, September 2020. ISSN 1464-8849. doi: 10.1177/1464884917738412. URL <https://doi.org/10.1177/1464884917738412>.
- José Ángel Gascón. Argumentative Bullshit. *Informal Logic*, 41(3):289–308, September 2021. ISSN 2293-734X, 0824-2577. doi: 10.22329/il.v41i3.6838. URL <https://dx.doi.org/10.22329/il.v41i3.6838>.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, November 2021. ISSN 0001-0782. doi: 10.1145/3458723. URL <https://doi.org/10.1145/3458723>.

- Jonathan Ginzburg. An update semantics for dialogue. In *Proceedings of the 1st International Workshop on Computational Semantics, Tilburg: ITK, Tilburg University, 1994*.
- Jonathan Ginzburg. Dynamics and the semantics of dialogue. *Logic, language and computation*, 1: 221–237, 1996.
- Jonathan Ginzburg. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, January 2012. ISBN 978-0-19-163248-8.
- Jonathan Ginzburg, Zulipiye Yusupujang, Chuyuan Li, Kexin Ren, Aleksandra Kucharska, and Pawel Lupkowski. Characterizing the Response Space of Questions: Data and theory. *Dialogue & Discourse*, 13(2):79–132, December 2022. ISSN 2152-9620. doi: 20221220143609000. URL <https://ojs3-prod.lib.uic.edu/ojs/index.php/dad/article/view/11531>.
- David Graeber. On the Phenomenon of Bullshit Jobs: A Work Rant. *Strike Magazine*, 3:1–5, August 2013. URL <https://www.strike.coop/bullshit-jobs/>.
- David Greatbatch. Aspects of Topical Organization in News Interviews: The Use of Agenda-Shifting Procedures by Interviewees. *Media, Culture & Society*, 8(4):441–455, October 1986. ISSN 0163-4437, 1460-3675. doi: 10.1177/0163443786008004005. URL <http://journals.sagepub.com/doi/10.1177/0163443786008004005>.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, February 2022. ISSN 2307-387X. doi: 10.1162/tacl\_a\_00454. URL [https://doi.org/10.1162/tacl\\_a\\_00454](https://doi.org/10.1162/tacl_a_00454).
- Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting Check-worthy Factual Claims in Presidential Debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, pages 1835–1838, Melbourne, Australia, 2015. ACM Press. ISBN 978-1-4503-3794-6. doi: 10.1145/2806416.2806652. URL <http://dl.acm.org/citation.cfm?doid=2806416.2806652>.
- Michael Heilman and Noah A. Smith. Good question! Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, pages 609–617, USA, June 2010. Association for Computational Linguistics. ISBN 978-1-932432-65-7.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/N13-1132>.
- Viktor Ivanković. Steering Clear of Bullshit? The Problem of Obscurantism. *Philosophia*, 44(2): 531–546, June 2016. ISSN 0048-3893, 1574-9274. doi: 10.1007/s11406-016-9709-8. URL <https://dx.doi.org/10.1007/s11406-016-9709-8>.

- Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering, January 2021. URL <http://arxiv.org/abs/2101.00391>.
- Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice*, 2(2):1–16, June 2021. ISSN 2692-1626, 2576-5337. doi: 10.1145/3412869. URL <https://dl.acm.org/doi/10.1145/3412869>.
- Jan Van Kuppevelt. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1): 109–147, March 1995. ISSN 1469-7742, 0022-2267. doi: 10.1017/S002222670000058X. URL <https://www.cambridge.org/core/journals/journal-of-linguistics/article/abs/discourse-structure-topicality-and-questioning/60F3E68601517091AF560CB3CC02C6AC>.
- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204, March 2020. ISSN 1560-4306. doi: 10.1007/s40593-019-00186-y. URL <https://doi.org/10.1007/s40593-019-00186-y>.
- Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL <https://aclanthology.org/C14-1141>.
- Marco Lippi and Paolo Torroni. Context-Independent Claim Detection for Argument Mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, June 2015. URL <https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/10942>.
- Shane Littrell, Evan F. Risko, and Jonathan A. Fugelsang. The Bullshitting Frequency Scale: Development and psychometric properties. *British Journal of Social Psychology*, 60(1):248–270, January 2021a. ISSN 0144-6665, 2044-8309. doi: 10.1111/bjso.12379. URL <https://onlinelibrary.wiley.com/doi/10.1111/bjso.12379>.
- Shane Littrell, Evan F. Risko, and Jonathan A. Fugelsang. ‘You can’t bullshit a bullshitter’ (or can you?): Bullshitting frequency predicts receptivity to various types of misleading information. *British Journal of Social Psychology*, page bjso.12447, February 2021b. ISSN 0144-6665, 2044-8309. doi: <https://dx.doi.org/10.1111/bjso.12447>. URL <https://dx.doi.org/10.1111/bjso.12447>.
- Paweł Łupkowski and Jonathan Ginzburg. Query responses. *Journal of Language Modelling*, 4(2): 245–292, 2016.
- MATRODA [@matrodamusic]. Corona Virus is Temporary. House music is forever, March 2020. URL <https://twitter.com/matrodamusic/status/1235617743161815041>.

- Daniel Mears. The ubiquity, functions, and contexts of bullshitting. *Journal of Mundane Behavior*, 3(2):233–256, June 2002. URL [https://www.researchgate.net/publication/289724155\\_The\\_ubiquity\\_functions\\_and\\_contexts\\_of\\_bullshitting](https://www.researchgate.net/publication/289724155_The_ubiquity_functions_and_contexts_of_bullshitting).
- Jörg Meibauer. Aspects of a theory of bullshit. *Pragmatics & Cognition*, 23(1):68–91, September 2016. ISSN 0929-0907, 1569-9943. doi: 10.1075/pc.23.1.04mei. URL <http://www.jbe-platform.com/content/journals/10.1075/pc.23.1.04mei>.
- Jörg Meibauer. The Linguistics of Lying. *Annual Review of Linguistics*, 4(1):357–375, January 2018. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-011817-045634. URL <http://www.annualreviews.org/doi/10.1146/annurev-linguistics-011817-045634>.
- Jörg Meibauer, editor. *The Oxford Handbook of Lying*. Oxford Handbooks in Linguistics. Oxford University Press, Oxford, United Kingdom, first edition edition, 2019. ISBN 978-0-19-873657-8. URL <https://global.oup.com/academic/product/the-oxford-handbook-of-lying-9780198736578?cc=de&lang=en&>.
- Jörg Meibauer. *Sprache und Bullshit*. Universitätsverlag WINTER Heidelberg, 2020. ISBN 978-3-8253-4808-3. URL [https://www.winter-verlag.de/de/detail/978-3-8253-4808-3/Meibauer\\_Sprache\\_und\\_Bullshit/](https://www.winter-verlag.de/de/detail/978-3-8253-4808-3/Meibauer_Sprache_und_Bullshit/).
- Chandra Mukerji. Bullshitting: Road Lore Among Hitchhikers\*. *Social Problems*, 25(3):241–252, February 1978. ISSN 0037-7791. doi: 10.2307/800062. URL <https://dx.doi.org/10.2307/800062>.
- Arvind Narayanan and Sayash Kapoor. ChatGPT is a bullshit generator. But it can still be amazingly useful, December 2022. URL <https://aisnakeoil.substack.com/p/chatgpt-is-a-bullshit-generator-but>.
- Condé Nast. ChatGPT’s Fluent BS Is Compelling Because Everything Is Fluent BS. *Wired UK*, 2022. ISSN 1357-0978. URL <https://www.wired.co.uk/article/chatgpt-fluent-bs>.
- OpenAI. ChatGPT: Optimizing Language Models for Dialogue, November 2022. URL <https://openai.com/blog/chatgpt/>.
- Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844, December 2019. doi: 10.1109/ASRU46091.2019.9003958.
- Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2018. ISSN 00100277. doi: 10.1016/j.cognition.2018.06.011. URL <https://dx.doi.org/10.1016/j.cognition.2018.06.011>.
- Gordon Pennycook and David G. Rand. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88(2):185–200, April

2020. ISSN 0022-3506, 1467-6494. doi: 10.1111/jopy.12476. URL <https://dx.doi.org/10.1111/jopy.12476>.
- Gordon Pennycook and David G. Rand. The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5):388–402, May 2021. ISSN 13646613. doi: 10.1016/j.tics.2021.02.007. URL <https://dx.doi.org/10.1016/j.tics.2021.02.007>.
- Gordon Pennycook, James Allan Cheyne, Nathaniel Barr, Jonathan A Fugelsang, and Derek J Koehler. On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, 10(6):549–563, November 2015. URL <https://psycnet.apa.org/record/2015-54494-003>.
- John V. Petrocelli. *The Life-Changing Science of Detecting Bullshit*. St. Martin’s Press, New York, first edition edition, 2021. ISBN 978-1-250-27162-4 978-1-250-28015-2. URL <https://us.macmillan.com/books/9781250271624/thelifechangingscienceofdetectingbullshit>.
- Barbara Plank. The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.731>.
- Livia Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5):601–638, December 1988. ISSN 0378-2166. doi: 10.1016/0378-2166(88)90050-1. URL <https://www.sciencedirect.com/science/article/pii/0378216688900501>.
- Dimas Wibisono Prakoso, Asad Abdi, and Chintan Amrit. Short text similarity measurement methods: A review. *Soft Computing*, 25(6):4699–4723, March 2021. ISSN 1433-7479. doi: 10.1007/s00500-020-05479-2. URL <https://doi.org/10.1007/s00500-020-05479-2>.
- Didik Dwi Prasetya, Aji Prasetya Wibawa, and Tsukasa Hirashima. The performance of text similarity algorithms. *International Journal of Advances in Intelligent Informatics*, 4(1):63, March 2018. ISSN 2548-3161, 2442-6571. doi: 10.26555/ijain.v4i1.152. URL <http://ijain.org/index.php/IJAIN/article/view/152>.
- Daniele Quercia, Harry Askham, and Jon Crowcroft. TweetLDA: Supervised topic classification and link prediction in Twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci ’12, pages 247–250, New York, NY, USA, June 2012. Association for Computing Machinery. ISBN 978-1-4503-1228-8. doi: 10.1145/2380718.2380750. URL <https://doi.org/10.1145/2380718.2380750>.
- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT\* ’20, pages 33–44, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372873. URL <https://doi.org/10.1145/3351095.3372873>.

- Florian Richter, Philipp Koch, Oliver Franke, Jakob Kraus, Fabrizio Kuruc, Anja Thiem, Judith Högerl, Stella Heine, and Konstantin Schöps. Open discourse, 2020. URL <https://doi.org/10.7910/DVN/FIKIBO>.
- Arndt Riester. Constructing QUD Trees. In Malte Zimmermann, Klaus von Heusinger, and V. Edgar Onea Gaspar, editors, *Questions in Discourse*, chapter Questions in Discourse, pages 164–193. Brill, March 2019. ISBN 978-90-04-37832-2. doi: 10.1163/9789004378322\_007. URL <https://brill.com/view/book/edcoll/9789004378322/BP000006.xml>.
- Arndt Riester, Lisa Brunetti, and Kordula Kuthy. Annotation Guidelines for Questions under Discussion and Information Structure. *HAL*, pages 1–56, 2018. doi: 10.1075/slcs.199.14rie. URL <https://hal.archives-ouvertes.fr/hal-01794160>.
- Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:1–69, December 2012. ISSN 1937-8912. doi: 10.3765/sp.5.6. URL <http://semprag.org/article/view/sp.5.6>.
- Carl Sagan. The Fine Art of Baloney Detection, 1996. URL <https://fermatlibrary.com/s/the-fine-art-of-baloney-detection>.
- John R. Searle. A Classification of Illocutionary Acts. *Language in Society*, 5(1):1–23, 1976. ISSN 0047-4045. doi: 10.1017/S0047404500006837. URL <https://www.jstor.org/stable/4166848>.
- Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. Overview of the CLEF-2021 Check-That! Lab: Task 3 on Fake News Detection. In *CEUR Workshop Proceedings*, volume 2936, page 18, Bucharest, Romania, 2021. CEUR-WS. URL <http://ceur-ws.org/Vol-2936/paper-30.pdf>.
- Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, page 12, 2021. doi: 10.18653/v1/2021.eacl-main.147. URL <https://dx.doi.org/10.18653/v1/2021.eacl-main.147>.
- André Spicer. Playing the Bullshit Game: How Empty and Misleading Communication Takes Over Organizations. *Organization Theory*, 1:1–26, 2020. doi: 10.1177/2631787720929704. URL <https://dx.doi.org/10.1177/2631787720929704>.
- Andreas Stokke and Don Fallis. Bullshitting, Lying, and Indifference toward Truth. *Ergo, an Open Access Journal of Philosophy*, 4(10):277–309, June 2017. ISSN 2330-4014. doi: 10.3998/ergo.12405314.0004.010. URL <https://dx.doi.org/10.3998/ergo.12405314.0004.010>.
- Donald Trump. Donald Trump’s News Conference: Full Transcript and Video, January 2017. URL <https://www.nytimes.com/2017/01/11/us/politics/trump-press-conference-transcript.html>.
- Donald Trump. Trump: ‘I’m the least racist person anybody is going to meet’ - BBC News, January 2018. URL <https://www.bbc.com/news/av/uk-42830165>.

- Christiane von Stutterheim and Wolfgang Klein. Referential Movement in Descriptive and Narrative Discourse. In RAINER Dietrich and CARL F. Graumann, editors, *North-Holland Linguistic Series: Linguistic Variations*, volume 54 of *Language Processing in Social Context*, pages 39–76. Elsevier, January 1989. doi: 10.1016/B978-0-444-87144-2.50005-7. URL <https://www.sciencedirect.com/science/article/pii/B9780444871442500057>.
- Sida Wang and Christopher D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL ’12, pages 90–94, USA, July 2012. Association for Computational Linguistics.
- Matthijs Westera, Laia Mayol, and Hannah Rohde. TED-Q: TED Talks and the Questions they Evoke. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1118–1127, Marseille, 2020. European Language Resources Association (ELRA). URL <https://aclanthology.org/2020.lrec-1.141/>.
- Xiaojing Yu and Anxiao Jiang. Expanding, Retrieving and Infilling: Diversifying Cross-Domain Question Generation with Flexible Templates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, page 11, 2021. doi: 10.18653/v1/2021.eacl-main.279. URL <https://dx.doi.org/10.18653/v1/2021.eacl-main.279>.
- Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438, 2021. ISSN 1749-818X. doi: 10.1111/lnc3.12438. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12438>.
- Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. Asking Too Much? The Rhetorical Role of Questions in Political Discourse. *arXiv:1708.02254 [physics]*, August 2017. URL <http://arxiv.org/abs/1708.02254>.
- Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural Question Generation from Text: A Preliminary Study. In Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, pages 662–671, Cham, 2018. Springer International Publishing. ISBN 978-3-319-73618-1. doi: 10.1007/978-3-319-73618-1\_56.
- Xinyi Zhou and Reza Zafarani. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5):1–40, October 2020. ISSN 0360-0300, 1557-7341. doi: 10.1145/3395046. URL <http://arxiv.org/abs/1812.00315>.