Issues in Information Systems
Volume 26, Issue 1, pp. 154-170, 2025

DOI: https://doi.org/10.48009/1_iis_112

A fake news detection framework

Seung C. Lee, *University of Minnesota Duluth, slee@d.umn.edu*

Abstract

In today's digital world, social media accelerates the spread of information, offering both advantages and serious risks. While it promotes open access to knowledge, it also enables the rapid circulation of fake news, misinformation, and disinformation. Fake news—deliberately false content designed to mimic legitimate media—threatens democratic values and public trust. Misinformation, shared without harmful intent, and disinformation, spread deliberately to mislead, further complicate the issue. Advanced AI technologies have made it harder to separate truth from falsehood. Combating fake news demands more than technical detection tools; it also requires promoting digital literacy and critical thinking. While current research often focuses on algorithmic and statistical detection, less attention has been given to the role of reasoning. This study proposes a new framework using counterfactual reasoning—imagining alternative scenarios—to evaluate information credibility. It also addresses modal fallacies, which occur when people confuse what is possible with what is necessary. Counterfactuals can reveal inconsistencies in fake news, but if used without care, they may lead to faulty logic. By identifying and correcting these reasoning errors, this approach enhances the detection of deceptive content and encourages more thoughtful information analysis.

Keywords: fake news, misinformation, disinformation, fake news detection, named entity recognition, relationship extraction, counterfactual reasoning, modal fallacies

Introduction

In the digital age, the rapid dissemination of information through social media platforms offers both benefits and challenges. On one hand, it democratizes information sharing, enabling people to disseminate news and ideas beyond the confines of traditional media (Miranda and Saunders, 2003). On the other hand, it facilitates the spread of fake news, misinformation, and disinformation, which can significantly affect public opinion, political processes, and public health (Allcott & Gentzkow, 2017).

Fake news is intentionally created to be sensational, misleading, or completely false, mimicking the appearance of legitimate news without adhering to similar ethical standards or editorial processes (Lazer et al., 2018). It is distributed across various channels, including social media, with the aim of deceiving or manipulating the audience (Pennycook & Rand, 2019). This overlaps with other forms of problematic information, such as misinformation and disinformation (Lazer et al., 2018). Misinformation is inaccurately spread without malicious intent, often due to errors or misunderstandings. It can stem from reporting inaccuracies, misinterpretation of data, or rumors, potentially causing harm by misleading the public (Pennycook et al., 2020 & Vosoughi et al., 2018). Disinformation, however, is false information spread deliberately to deceive. It can manifest in various deceptive forms, aiming to influence opinions, erode trust, or promote specific agendas (Fallis, 2015, Phillips, 2015).

Worse, the rise of advanced artificial intelligence (AI) technologies complicates the identification of authentic content, as these algorithms can generate realistic text, images, and videos (Corse, 2024 & Murphy, 2024). The use of such technology for malicious purposes, such as creating convincing fake news or deepfakes, poses a significant challenge in distinguishing fact from fiction. The consequences of fake news can be severe, undermining democracy, polarizing societies, and even causing public health crises (Fallis, 2015 & Kshetri and Voas, 2017).

Addressing fake news requires a comprehensive approach that includes developing better detection algorithms, raising public awareness, enhancing digital literacy, and fostering critical thinking (Wang et al., 2022). The integrity of democratic processes and public trust depends on effectively managing the spread of fake news (Wardle & Derakhshan, 2017). Research on fake news detection has focused on its false knowledge, writing style, propagation patterns, and source credibility (Zhou and Zafarani, 2020). However, there is a need for methods that are accessible to everyday users, such as counterfactual reasoning and avoiding modal fallacies, which can help in identifying misinformation.

This study introduces a novel framework for detecting fake news using counterfactual reasoning, which involves imagining alternative scenarios to assess the credibility of information. This approach can help evaluate news sources, assess story consistency, identify manipulated content, and encourage critical thinking (Stanford Encyclopedia of Philosophy, 2019). However, relying solely on counterfactual reasoning can introduce challenges, such as confusion over possibility and necessity, making it crucial to avoid modal fallacies (Hughes and Cresswell, 1996 & Walton, 1992). Identifying such fallacies can significantly improve the detection of fake news. The paper concludes with a discussion on the effectiveness of this framework.

Theoretical Background

Counterfactual Theories of Causation

Detection of fake news entails the examination of a confluence of three interrelated parameters: antecedent, decision, and consequent. The decision to accept fake news as true news or refute it as a whole or partially by an individual user can be articulated as a factual proposition comprising two distinct segments: antecedent and consequent. Consider the factuals below:

- Increasing access to education leads to higher levels of economic prosperity.
- An online trading company specializing in binary options, cryptocurrency, and forex trading guaranteed higher returns using the latest technology, but I ended up losing most of my investment.

In logic, the antecedent is the first part of a proposition, and the consequent follows; in fake news detection, the truth of the consequent often depends on the credibility of the antecedent. Users' evaluations are also shaped by cognitive and behavioral biases that may not be apparent to researchers (Hu et al., 2022). Establishing causality between antecedent and consequent aligns with counterfactual theories of causation, which consider what would happen if a key event had not occurred. David Hume famously described this in 1748, noting that a cause is something without which its effect would not exist (Millican, 2007). In philosophy and related fields, causal reasoning is often expressed through counterfactual conditionals such as "If A had not occurred, C would not have occurred" (Stanford Encyclopedia of Philosophy, 2019). Table 1 illustrates a set of such counterfactuals, demonstrating how hypothetical alternatives can help assess the plausibility of causal claims within a narrative.

Factual	Counterfactual
Increasing access to education leads to higher levels of economic prosperity.	If access to education had not been increased, economic prosperity would have remained stagnant.
	Without increased access to education, economic prosperity would not have seen significant growth.
	Had there been no increase in access to education, economic prosperity would have been limited.
	If access to education had not improved, economic prosperity would have suffered.
	If there had been no efforts to increase access to education, economic prosperity would not have flourished.
An online trading company specializing in binary options, cryptocurrency, and forex trading guaranteed higher returns using the latest technology, but I ended up losing most of my investment.	If the online trading company had not guaranteed higher returns, I would have approached my investment with more caution, potentially avoiding significant losses.
	If the latest technology had not been a feature of the company, my confidence in their ability to deliver on promises would have been lower, possibly leading me to invest less or not at all.
	If the company had not specialized in high-risk areas like binary options, cryptocurrency, and forex trading, I would have chosen a different, potentially safer, investment path.
	If the technology and strategies promised by the company had actually been effective in securing higher returns, I would not have experienced a significant loss on my investment.
	If I had been more skeptical of the promises made by the company and sought independent advice, I could have avoided or minimized my investment, thereby reducing the risk of loss.

Counterfactuals have been widely used in decision-making by comparing actual and hypothetical scenarios to guide judgments (Cantone, 2020; Zhang et al., 2022). These analyses offer a structured approach to preventing future harms and are often formulated using methods like structural equations (Woodward & Hitchcock, 2003). Among multiple counterfactuals for a given event, some can be optimized to maximize outcome utility—these are known as optimized counterfactuals and align with rational choice theory (Marwala, 2014). However, real-world decision-making is constrained by limited information, cognitive biases, and time, making such optimization bounded. This aligns with the theory of bounded rationality (Marwala & Hurwitz, 2017).

At its core, counterfactual theory clarifies causality using conditional reasoning—expressed as "If A had not occurred, C would not have occurred" (Menzies & Beebee, 2019). This reasoning is crucial in fake news detection, influencing users' interpretation and belief in disinformation. According to Hempel and Oppenheim's (1948) deductive-nomological (D-N) model, valid explanations require that the explanandum logically follows from true premises in the explanans, one of which must reflect a regular law. This framework supports causal reasoning in detecting and explaining fake news. For instance, actions like responsible sharing (Pennycook & Rand, 2019) and critical engagement (Walter et al., 2018) can be seen as explanatory mechanisms for reducing misinformation's spread.

Generalizations that fit the deductive-nomological (D-N) model—or exhibit invariance even without being natural laws—can still support counterfactual reasoning in hypothetical scenarios (Woodward, 1996;

Woodward & Hitchcock, 2003). These counterfactual conditionals show how the consequent would change under various interventions on the antecedent, illustrating causality beyond empirical laws (Marwala, 2014). At their core, counterfactual theories posit that causal relationships are best understood through conditionals like "If A had not occurred, C would not have occurred" (Stanford Encyclopedia of Philosophy, 2019). This framework is particularly relevant to fake news detection, where evaluating the link between cause and effect shapes user judgments. According to Hempel and Oppenheim (1948), a valid explanation requires that the explanandum logically follow from true premises (explanans), one of which must reflect a general law or regularity.

Lewis (1973, 1986) developed the most influential counterfactual theory of causation, proposing that causal dependence exists when the absence of one event would prevent another. Using the semantics of possible worlds, he argued that a counterfactual is true if the closest possible world where the antecedent holds also contains the consequent. This comparative similarity framework defines causality in terms of deviations from actuality. Lewis's notion of causal dependence rests on three conditions: the events must be distinct, involve the right (non-backtracking) counterfactuals, and be considered as events rather than general facts (Mellor, 1995, 2004). Although Lewis's theory remains foundational, subsequent critiques (Elga, 2000; Hall, 2004; Paul & Hall, 2013) highlight its limitations for analyzing singular causation. In response, structural equation models have emerged as the dominant alternative (Hitchcock, 2001, 2007; Woodward, 2003). These models represent causal relationships as an ordered pair ⟨V, E⟩, where E is a set of equations connecting variables in V that denote potential states of a system. Variables can be binary (e.g., 1 for event occurrence, 0 for absence) or multi-valued, and include both exogenous and endogenous elements. This framework enables precise, testable representations of causality and has become central to contemporary counterfactual analysis.

Each equation in *E* specifies the value of a single variable appearing on the left-hand side of the equation, with exactly one equation per variable. *E* consists of two subsets: one subset containing equations with exogenous variables on the left-hand side and another subset containing equations with endogenous variables on the left-hand side. Equations in the former subset typically adopt a simple form, such as $Z = z$, indicating the actual value of the variable in question. For instance, equations describing attitudes toward mask-wearing across cultures fall within this subset. Equations in the latter subset express the value of the endogenous variable as a function of the values of other variables in set *V*, taking the form:

$$\mathbf{Y = f(X_1,\ldots, X_n).}$$

For example, *Y* might represent the number of individuals contracting coronavirus, with X_n denoting various preventive measures like maintaining social distance, wearing masks, avoiding prolonged person-to-person interactions, practicing hand hygiene, the intensity of preventive efforts, the presence of non-compliant individuals, and so forth. Though interpretations of this structural equation framework vary (Pearl, 2000), Woodward (2003) and Hitchcock (2001) view it as expressing fundamental counterfactuals of the form:

If it were the case that $X_1 = x_1$, $X_2 = x_2$,…, $X_n = x_n$, then it would be the case that $Y = f(x_1,\ldots,x_n)$.

This type of counterfactual implies reading structural equations from right to left: the antecedent specifies potential values of variables X_1 through X_n, while the consequent indicates the corresponding value of endogenous variable *Y*. Such counterfactuals exist for every combination of potential values of variables X_1 through X_n. Moreover, this structural equation type inherently embodies a right-to-left asymmetry, mirroring the asymmetry of non-backtracking counterfactuals. For instance, in a factual scenario where Jane does not wear a mask and contracts coronavirus, the non-backtracking counterfactual "If Jane had

worn a mask, she would not have contracted influenza" holds true. Conversely, the counterfactual "If Jane had contracted influenza, she would have worn a mask" is false.

Modal Logic

Modal logic is an extension of classical logic that introduces modalities, which express notions of possibility and necessity. In modal logic, statements can be qualified not only in terms of their truth or falsity but also in terms of whether they are necessarily true, possibly true, or neither. The basic modalities are often represented by the operators ◇ (diamond) for possibility and □ (box) for necessity (Hardegree, 2009 & Modal Logic, 2023). Necessity (□) means that a statement is necessarily true if it is true in all possible worlds or under all possible circumstances. For example,

$$\Box(2+2=4)$$

could represent the statement "It is necessarily true that 2+2 equals 4," indicating that in any conceivable scenario, 2+2 will always equal 4. Meanwhile, possibility (◇) means that a statement is possibly true if there is at least one possible world or circumstance in which it is true. For example,

$$\Diamond(\text{There is life on other planets})$$

could signify "It is possible that there is life on other planets," meaning there's at least one conceivable scenario where this statement holds true. Relationships between modalities can also be defined. For example, the statement ◇P is equivalent to ¬□¬P, which means "P is possible" is equivalent to saying, "It is not necessarily the case that P is not true." Conversely, □P is equivalent to ¬◇¬P, meaning "P is necessary" is equivalent to saying, "It is not possible that P is not true."

Let's take an example of necessity. Consider:

Statement: "All bachelors are unmarried."
Modal Logic: $\Box(Bachelor(x) \rightarrow Unmarried(x))$

The modal logic is interpreted that it is necessarily true that if someone is a bachelor, then they are unmarried. This necessity holds under all conceivable conditions. Now let's take an example of possibility. Consider:

Statement: "It's possible for it to rain tomorrow."
Modal Logic: $\Diamond(Rain(Tomorrow))$

Modal logic is interpreted to mean that there exists at least one scenario or possible world where it rains tomorrow. Modal logic is a versatile tool applied across disciplines to formalize reasoning about possibility and necessity. In philosophy, it provides a framework for analyzing metaphysical concepts such as existence and potentiality. In computer science, it supports program verification and system specification by defining conditions under which statements must necessarily or possibly hold, thus improving system reliability.

In linguistics, modal logic aids in analyzing natural language semantics, especially modal expressions conveying possibility and necessity. Its ability to go beyond the binary constraints of classical logic enables more nuanced analysis of meaning and truth conditions. By capturing the dynamics of modal concepts, modal logic plays a foundational role in both theoretical and applied domains of logic and reasoning.

Modal Fallacy

The modal fallacy represents a distinctive form of fallacious reasoning within modal logic, characterized by the misplacement of a proposition within an incorrect modal scope. This fallacy frequently manifests as a conflation of the scope pertaining to what is necessarily true. A proposition is deemed necessarily true if it is impossible for it to be false, and there exists no conceivable circumstance wherein the proposition could be negated. Certain philosophers extend this notion, contending that a necessarily true statement must hold true across all conceivable worlds (Routley, 1969).

Within the framework of modal logic, a proposition, denoted as P, may possess the property of being necessarily true or false, symbolized as □ P and □ ¬ P respectively. This denotes that the truth or falsity of the proposition is *logically predetermined*; alternatively, it may be possibly true or false, represented as ◇ P and ◇ ¬ P, indicating that while the proposition may *indeed be true or false*, its veracity is not logically predetermined but contingent upon circumstances. The modal fallacy arises when there is a failure to discern the distinction between these modal categories. Within modal logic, a crucial distinction exists between what is logically necessary to be true and what is true but not logically necessary. One common manifestation of this fallacy is the substitution of $p \rightarrow q$ with $p \rightarrow \Box q$. In the former, q is true given p but is not logically necessary to be so. That is, modal fallacies occur when there's an error in reasoning about necessity, possibility, or impossibility. The following are examples of affirming the denying the consequents.

Proposition: "If it's raining, then it's necessary that the streets are wet."
Modal Fallacy: Affirming the consequent.

This statement commits the fallacy of affirming the consequent. Just because the streets are wet doesn't necessarily mean it is raining. There could be other reasons for the streets being wet, such as someone watering their lawn or a nearby river overflowing. Therefore, concluding that the rain is the necessary cause of wet streets is fallacious. This mistake conflates correlation with causation and ignores other potential causes.

Proposition: "If John is in Paris, then he must be in France."
Modal Fallacy: Denying the consequent.

This statement commits the fallacy of denying the consequent. It wrongly concludes that if John is not in France, then he cannot be in Paris. However, this overlooks the possibility that there may be other locations within France where John could be besides Paris. Just because John is not in France doesn't necessarily mean he cannot be in Paris, as Paris is a city within France, but he could also be in other places within the country. Therefore, denying the consequent in this context leads to an erroneous conclusion by disregarding alternative possibilities.

The following are examples of affirming and denying the antecedents.

Proposition: "If I were a millionaire, then I would be happy."
Modal Fallacy: Denying the antecedent.

This statement commits the fallacy of denying the antecedent. It assumes that being a millionaire is a necessary condition for happiness and implies that if someone is not a millionaire, they cannot be happy. However, this overlooks the possibility that there are many other factors besides wealth that contribute to happiness, such as relationships, health, personal fulfillment, etc. Therefore, just because someone is not a

millionaire doesn't mean they cannot be happy. This fallacy erroneously denies other potential sources of happiness besides wealth.

Proposition: "If it's snowing, then it must be cold."
Modal Fallacy: Affirming the antecedent.

This statement affirms the antecedent by concluding that if it's cold, then it must be snowing. However, this overlooks the possibility that there could be other reasons for it to be cold besides snowfall. For instance, it could be cold due to clear skies and a drop in temperature overnight. Therefore, just because it's cold doesn't necessarily mean it is snowing. Affirming the antecedent in this context leads to an erroneous conclusion by neglecting alternative explanations for the observed condition.

Counterfactual Theories of Causation and Modal Fallacy

Counterfactual theories of causation suggest that an event A causes an event B if, and only if, had A not occurred, B would not have occurred. This approach relies heavily on counterfactual conditionals ("If A had not happened, then B would not have happened") to analyze causal relationships. However, when evaluating these theories from the perspective of modal fallacy, several issues arise (Buckley, 1988, Piaget, 1986 & Stanford Encyclopedia of Philosophy, 2019).

Confusion between possibility and necessity

A modal fallacy occurs when there is a confusion between what is possible and what is necessary. In the context of counterfactual theories, this might mean erroneously assuming that because an event could cause another in some possible worlds, it necessarily causes it in the actual world. This overlooks the fact that causation requires more than just a possible connection; it requires a necessary connection within the actual world's conditions.

Difficulty in determining relevant possible worlds

Counterfactuals rely on comparing the actual world to a closest possible world where the supposed cause did not occur. Determining which possible world is the closest involves subjective judgments about which aspects of the world are held fixed and which can vary. This subjectivity can introduce modal fallacies by assuming some worlds are relevant to the causal analysis when they might not be, leading to incorrect conclusions about causation.

Overlooking overdetermination and preemption

Counterfactual theories can struggle with cases of overdetermination and preemption, where an event is caused by multiple factors or where one potential cause preempts another. In such cases, removing one cause might not prevent the effect because other causes could still bring it about, suggesting a modal fallacy by implying a necessary connection between a specific cause and an effect when multiple causal pathways exist.

Assumption of causal determinism

The use of counterfactuals often implicitly assumes a form of causal determinism, where given conditions necessarily lead to specific outcomes. This assumption can be challenged by quantum mechanics and other indeterministic theories, suggesting a modal fallacy by conflating the actual indeterministic nature of the world with a deterministic interpretation of causality.

Misapplication of counterfactuals to complex systems

In complex systems, small changes can lead to vastly different outcomes due to chaotic dynamics. Applying counterfactual reasoning in these contexts might lead to modal fallacies by assuming linear and predictable causal relationships in systems where sensitivity to initial conditions renders such simplicity and

predictability moot. Addressing these issues requires careful attention to the distinctions between necessity and possibility, a nuanced approach to selecting relevant possible worlds, and an acknowledgment of the complexities inherent in causal analysis, especially in indeterministic or highly complex systems.

A Fake News Detection Framework

The concept of a counterfactual cluster suggests that for any fake news instance, multiple counterfactuals can be generated based on various combinations of antecedent variables. Each piece of fake news can thus be decomposed into antecedents and consequents and reframed as a cluster of rational counterfactuals. This allows for a broader detection framework that accounts for both explicit and implicit variables. The approach aligns with Lewis's semantics of possible and actual worlds (1973) and Hitchcock's structural equations framework for causation (2001). However, relying solely on counterfactual reasoning raises challenges, highlighting the need to also address modal fallacies. A robust fake news detection framework should integrate counterfactual reasoning and modal logic to evaluate not only the factual accuracy but also the causal and logical coherence of information. This requires the combined use of natural language processing, causal inference, and formal reasoning to assess the truthfulness and structure of news content. Figure 1 illustrates our fake news detection framework.

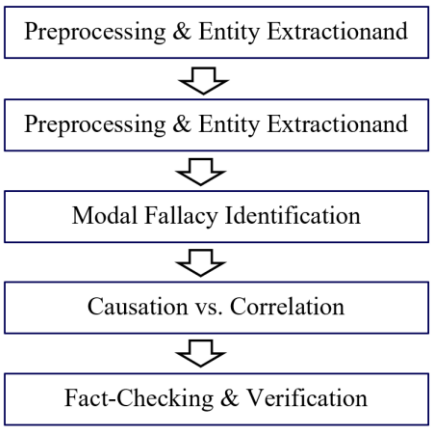

Figure 1. Fake news detection framework.

Preprocessing and Entity Extraction

The objective of this step is to extract and identify key variables, entities, and the causal relationships implied in the text. For this purpose, we can use natural language processing (NLP) techniques like Named Entity Recognition (NER) and Relation Extraction (RE) to parse the text and identify these elements. NER stands as a core technology within the realm of NLP. NER is the process of identifying and classifying key variables and entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. (Nadeau and Sekine, 2007). For example, consider the fake news or claim: "Disney World is going to lower the drinking age to 18. Disney World is battling the Florida government in court to get a resort exemption, which would allow anyone 18 and older to drink on property." NER would then identify the following:

- **Disney World** would be recognized as an **ORGANIZATION**.
- **18** would be recognized as a **CARDINAL.**
- **Florida government** would be recognized as an **ORGANIZATION**.

- **court** may or may not be recognized as an **ORGANIZATION**, depending on the context and the NER model's interpretation.
- **resort exemption** would be recognized as a **MISC**.
- **18 or older** would be recognized as a **CARDINAL**, with "older" possibly not being recognized as part of the entity.

Relationship extraction (RE) is a subfield of natural language processing (NLP) that focuses on identifying and categorizing the connections between entities mentioned in text (Bach and Badaskar, 2007). These entities can be anything from people and organizations to locations and dates. RE goes beyond simply recognizing entities; it attempts to understand the semantic relationships that link them together. That is, RE is the process of identifying and classifying semantic relationships between entities within a text. This task aims to find and categorize predefined relationships (such as "is employed by," "is a," "guarantees," "located in," etc.) between entities recognized by NER. For example, given the entities identified in the previous NER example, RE would work to identify relationships between them, such as:

- **Disney World** ORGANIZATION - [attempts_to_lower_drinking_age_to] -> CARDINAL **18**
- **Disney World** ORGANIZATION - [battles_in_court_with] -> ORGANIZATION **Florida government**
- **Disney World** ORGANIZATION - [seeks] -> MISC **resort exemption**
- **resort exemption** MISC - [allows_drinking_for] -> CARDINAL **18 or older**

The relationships here help in structuring the information extracted from the text, making it clearer how entities are related to each other within the context provided. This structured approach is beneficial for tasks such as fake news detection, where understanding the relationships between entities is crucial (De Magistris et al., 2022). In practice, NER and RE are often used together in information extraction systems to automatically extract structured information from unstructured texts (Nasar et al., 2021). This structured information can then be used for various applications, such as building knowledge graphs, feeding into databases, or facilitating better search and retrieval mechanisms. These tasks can be performed using NLP models that can understand the context and nuances of human language (Ni et al, 2022 & Yan et al., 2021).

Counterfactual Analysis

In this step, we evaluate the plausibility of the causal relationships by considering alternative scenarios (counterfactuals). To do so, we need to generate counterfactual scenarios and test if the causation holds under different conditions. For example, we could have the following counterfactual cluster for the output identified by NER + RE above:

- If Disney World had not attempted to lower the drinking age, it might have focused on enhancing family-friendly attractions or implementing new safety protocols instead.
- If Disney World had collaborated with the Florida government instead of battling in court, they could have jointly launched initiatives to promote responsible alcohol consumption among young adults.
- If Disney World had sought a sustainability exemption rather than a resort exemption for drinking, it would have prioritized environmental initiatives, like reducing waste or conserving water, to attract eco-conscious visitors.
- If Disney World had advocated for raising the drinking age to 21 instead of lowering it to 18, it might have positioned itself as a leader in promoting health and safety among its visitors.
- If the Florida government had proposed lowering the drinking age for resorts, aiming to boost tourism, Disney World could have been at the forefront of implementing and managing this policy change.

These counterfactual scenarios or, simply, counterfactuals illustrate how changing key aspects of the original statement could significantly alter the attempt's context, execution, and perceived importance.

Modal Fallacy Identification

In this step, we identify and evaluate the logical structure of the claims for modal fallacies. To do that, we analyze the language to detect modal verbs that are used to present possibilities as certainties. Modal fallacies involve errors in reasoning regarding necessity and possibility. Let's examine the statement for such fallacies: "If Disney World had not attempted to lower the drinking age, it would have focused on enhancing family-friendly attractions or implementing new safety protocols instead." The statement uses modal language ("might") to express a possibility without asserting necessity. Here's the breakdown:

- **Possibility vs. Necessity:** The use of "would have focused" implies a strong likelihood or certainty about the alternative actions Disney World might have taken if it had not attempted to lower the drinking age. However, from the perspective of possibility, this statement overlooks the range of possible actions Disney World could have chosen. It is indeed possible that Disney World might have chosen to enhance family-friendly attractions or implement new safety protocols as alternative strategies. However, it is not the only possible outcome; there are numerous other strategies Disney World could have considered to attract visitors or improve the guest experience without lowering the drinking age. The statement implies a necessity that these specific actions (enhancing family-friendly attractions or implementing new safety protocols) would be the direct and inevitable response to not attempting to lower the drinking age. This implication is not supported by a necessary causal link; the decision not to lower the drinking age does not logically necessitate that the only or inevitable alternative would be to focus on family-friendly attractions or safety protocols. The organization could have pursued a variety of other initiatives, such as marketing campaigns, ticket pricing strategies, or different types of guest experiences that are not directly related to the drinking age or even to safety and family-friendly attractions.

- **Causal relationship:** This statement implies a direct causal relationship between not attempting to lower the drinking age and choosing to focus on enhancing family-friendly attractions or implementing new safety protocols. The modal fallacy here could be identified as a false dilemma or false dichotomy, suggesting that Disney World's only alternative actions to not lowering the drinking age would be to enhance family-friendly attractions or implement new safety protocols, excluding other possible actions. Additionally, the statement could imply a form of determinism that is not warranted without further evidence, suggesting that one specific outcome (focusing on family-friendly attractions or safety protocols) would necessarily follow from the condition (not attempting to lower the drinking age), which may not account for the complexity of decision-making in a large organization.

In summary, the statement commits an implicit assumption of necessity where only possibility exists. It suggests a direct and necessary causal relationship where there should only be considered a range of possible outcomes. This overstatement of certainty fails to acknowledge the complex decision-making process within an organization like Disney World, where multiple factors influence strategic choices beyond the binary implied by the statement. Therefore, the statement oversimplifies the outcomes as necessary when they should be framed as one of many possibilities, highlighting the need for nuanced understanding of causal relationships in speculative scenarios.

Causation vs. Correlation Analysis

After identifying a modal fallacy, particularly one that involves assumptions about what must or might happen, it’s important to question whether the argument mistakenly conflates correlation with causation. Causation vs. correlation analysis is essential for accurately interpreting relationships between variables or

events. Understanding this distinction is crucial, especially after identifying modal fallacies, as it further clarifies the logical and empirical boundaries of claims or arguments.

Causation

After identifying a modal fallacy, if the fallacy involves assuming a necessary outcome (e.g., assuming that one action must inevitably lead to a specific result), evaluating for causation involves critically examining whether a direct cause-and-effect relationship actually exists, as claimed. Causation requires rigorous evidence, often involving controlled experiments or longitudinal studies that can isolate variables and demonstrate that the cause precedes the effect and that the relationship holds even when controlling for other potential causes.

Correlation

Correlation, on the other hand, refers to a statistical measure that describes the extent to which two variables change together, but it does not imply that one variable causes the change in the other. Correlations can be positive (both variables increase or decrease together), negative (one variable increases when the other decreases), or null (no relationship). Identifying correlations can be misleading if interpreted as causation without further analysis. For instance, if a modal fallacy involves assuming causation based on a mere correlation (e.g., two events or trends occurring simultaneously), critical analysis is needed to clarify that the observed relationship does not necessarily imply one causes the other. Correlation analysis involves looking at the data to see if there is a statistical relationship between variables but also acknowledging that this relationship could be due to chance, common causes, or other variables not considered in the analysis.

Using the statement "If Disney World had not attempted to lower the drinking age, it would have focused on enhancing family-friendly attractions or implementing new safety protocols instead." as a basis, let's explore examples of causation and correlation to illuminate the differences and common misunderstandings that can arise. The statement suggests a relationship between Disney World's decision-making regarding the drinking age and its focus on other initiatives like enhancing family-friendly attractions or safety protocols. For a correlated scenarion, suppose observation of data shows that in years when Disney World introduces more family-friendly attractions, there are fewer initiatives to change age-related policies.

Also suppose that there is a statistical correlation between the number of new family-friendly attractions and the lack of initiatives to change the drinking age. As the number of family-friendly attractions increases, efforts to change the drinking age decrease. This leads to correlation misinterpretation. One might misinterpret this correlation as indicating that focusing on family-friendly attractions causes Disney World to not pursue changes in the drinking age policy. However, this is just a correlation; both trends could be independently decided or influenced by other factors, such as broader company strategy, market research, or legal advice.

For example, we use the same statement. The implication of the statement is that the decision not to lower the drinking age directly causes Disney World to allocate resources to other areas, like enhancing attractions or safety. For a causal scenario, suppose Disney World explicitly states in a strategy meeting that due to the decision not to pursue lowering the drinking age, it will redirect funds and efforts towards enhancing family-friendly attractions and safety protocols. This decision implies causation that the decision not to lower the drinking age is the direct cause of increased investment in family-friendly attractions and safety protocols.

The causal link is established by the decision-making process, where resources are intentionally redirected from one initiative (changing the drinking age) to another (enhancements and safety). In this scenario, there's clear evidence of a cause-and-effect relationship. The decision-making process at Disney World (cause) directly leads to a specific redirection of focus and resources (effect). This is not merely a correlation

because there's an identifiable mechanism (decision and resource allocation) that links the cause with the effect. Correlation in this context might suggest that whenever Disney World focuses less on policy changes like the drinking age, it coincidentally increases its emphasis on family-friendly attractions and safety. This doesn't prove one action causes the other; they might both stem from a broader strategic focus or external pressures.

Causation implies that Disney World's decision-making process directly leads to specific outcomes. If Disney World decides not to pursue a change in the drinking age, and it is explicitly stated that as a result, resources are shifted towards enhancing attractions and safety protocols, this establishes a direct cause-and-effect relationship. Understanding the distinction between causation and correlation is crucial in evaluating statements, policies, and strategies to avoid misleading conclusions. In the case of Disney World's decision-making, scrutinizing the evidence for direct links and the presence of other influencing factors can clarify whether observed relationships are causal or merely correlative.

Fact-Checking and Verification

Fact-checking and verification are critical processes in the evaluation of claims, statements, and assertions made in various types of content. The objective is to ensure the accuracy and reliability of information before it is disseminated or used as a basis for decisions, opinions, or further research. The primary goal of fact-checking and verification is to confirm the truthfulness and accuracy of the information. This is crucial in maintaining the integrity of discourse. It helps prevent the spread of misinformation and disinformation, which can lead to misunderstandings, false beliefs, and potentially harmful decisions (Uscinski and Butler, 2013). The methodology for fact-checking involves several steps and resources:

1. **Identify the claim:** Clearly define what claim or statement needs verification. The claim should be specific enough to be verifiable.
2. **Source evaluation:** Determine the original source of the claim. Is it a primary source (e.g., a research study, an eyewitness account) or a secondary source (e.g., a news article summarizing a study)?
3. **Search for reputable sources:** Look for information from reliable, authoritative sources to confirm the claim. These sources include peer-reviewed journals, official reports, government databases, and recognized news organizations. The credibility of the source is paramount.
4. **Check multiple sources:** Verify the claim across several reputable sources to ensure consistency and reliability of the information. Different perspectives from multiple sources can provide a more comprehensive understanding.
5. **Evaluate the evidence**: Assess the quality and relevance of the evidence supporting the claim. For scientific claims, this might involve reviewing study methodologies, sample sizes, and the significance of the findings.
6. **Consider the context:** Understand the broader context of the claim. Sometimes, information may be factually correct but misleading if presented without sufficient context.

Let’s apply the method to the claim that "chocolate causes weight loss":

1. **Identify the claim:** The specific claim is that consuming chocolate leads to weight loss.
2. **Source evaluation:** Find the original source of the claim. Was it made in a scientific study, a news article, or a blog post?
3. Search for reputable sources: Look for nutritional studies in peer-reviewed journals or statements from health organizations about the effects of chocolate on weight loss.
4. **Check multiple sources:** Compare findings from different studies and articles. Are there meta-analyses or systematic reviews on the topic?

5. **Evaluate the evidence**: Review the methodologies of the studies supporting the claim. Are the study designs robust and the results statistically significant? Do the studies account for variables such as the type of chocolate, the amount consumed, and the dietary context?
6. **Consider the context:** Assess any conditions or limitations mentioned in the studies. For example, does chocolate contribute to weight loss only under certain conditions, such as in conjunction with exercise or a specific diet?

An Example Application of the Framework

Let's look at a simple application of the framework. Consider a piece of fake news stating, "A new study proves that drinking two cups of coffee daily cures Alzheimer's disease." The framework would:

1. **Extract facts:** Identify entities (coffee, Alzheimer's disease) and the causal claim (drinking coffee → cures → Alzheimer's disease).
2. **Counterfactual analysis**: Generate scenarios like "Individuals drinking coffee who also follow a healthy lifestyle" to examine if coffee alone is the cure.
3. **Modal fallacy identification:** Check if the article improperly presents the study's findings as a definite cure rather than a possible correlation.
4. **Causation vs. correlation:** Analyze data to see if coffee directly impacts Alzheimer's disease or if the observed effect is due to other factors.
5. **Fact-checking**: Cross-reference the claim with medical research databases to find if there's credible evidence supporting it.

Discussion

The proposed framework for detecting fake news, focusing on counterfactual reasoning, modal fallacy identification, and causation versus correlation analysis, addresses the nuanced task of identifying misinformation. However, it confronts several limitations and implementation challenges, particularly when considering the multi-modal nature of fake news. The complexity of language and context, inherent in textual information, is compounded when fake news is disseminated through multiple modes, such as images, videos, and social media platforms. This multi-modality introduces additional layers of complexity, as the framework must not only analyze text but also understand visual cues and how they interact with textual information to convey misleading messages.

The dynamic nature of news and the rapid evolution of content formats further complicate the detection of multi-modal fake news. Adapting to new topics, contexts, and media formats requires the framework to be incredibly flexible and continuously updated. Issues of bias and subjectivity are also magnified in a multi-modal context, as different cultures and communities may interpret visual and textual cues in diverse ways. Moreover, the availability and accessibility of reliable data for cross-referencing and fact-checking become even more challenging when dealing with images and videos, which may be manipulated or taken out of context to support false narratives.

Implementing a detection system capable of analyzing multi-modal fake news, which is beyond the scope of this study, demands significant algorithmic complexity and computational resources. The system must accurately process and interpret not just text but also visual and auditory information, increasing the risk of false positives and negatives. The interdisciplinary nature of the challenge requires expertise not only in computer science and linguistics but also in fields such as computer vision and media studies.

Future directions for addressing multi-modal fake news involve leveraging advanced machine learning techniques that can effectively process and analyze data from various sources and formats. Collaborating

with experts across multiple disciplines becomes even more crucial to develop algorithms capable of understanding the nuanced interplay between text, images, and videos. Engaging the public and incorporating crowdsourced verification methods can also enhance the detection of multi-modal misinformation by utilizing the collective scrutiny of diverse audiences. Ensuring ethical guidelines and transparency in the algorithms' operation is vital for maintaining public trust, especially when dealing with the complexities of multi-modal content. Addressing the challenges posed by multi-modal fake news is essential for advancing the detection framework, requiring innovative solutions and collaborative efforts to navigate the intricacies of misinformation in today's digital landscape.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2), 211-236.

Bach, N., & Badaskar, S. (2007). A review of relation extraction. *Literature review for Language and Statistics II*, 2, 1-15.

Buckley, J. J. (1988). Possibility and necessity in optimization. Fuzzy sets and systems, 25(1), 1-13.

Cantone, J. A. (2020). Counterfactual Thinking, Causation, and Covariation in Mock Juror Assessments of Negligence: Twenty-Five Years Later. *Psychological Reports*, 123(2), 371-394.

Corse, A. (2024, January 25). Taylor Swift's Fans Swarm X to Combat AI Fakes of Singer. *Wall Street Journal*. https://www.wsj.com/tech/taylor-swifts-fans-swarm-x-to-combat-ai-fakes-of-singer-cf928031

De Magistris, G., Russo, S., Roma, P., Starczewski, J. T., & Napoli, C. (2022). An explainable fake news detector based on named entity recognition and stance classification applied to covid-19. *Information*, 13(3), 137.

Elga, A. (2000). Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science*, 68(3), Supplement, 313–24.

Fallis, D. (2015). What is disinformation?. *Library trends*, 63(3), 401-426.

Hall, N. (2004). *Two Concepts of Causation*. In Collins, J., Hall, N., and Paul, L. A. (eds.), pp. 225–76, Causation and Counterfactuals, The MIT Press. DOI: https://doi.org/10.7551/mitpress/1752.001.0001

Hardegree, G. (2009). An Introduction to Modal Logic, UMass Amherst.

Hempel, C. G., and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175. DOI:10.1086/286983

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, 98(6), 273–99.

Hitchcock, C. (2007). Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review*, 116(4), 495–532

Hu, L., Chen, Z., Yin, Z. Z. J., & Nie, L. (2022). Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(11), 11141-11152.

Hughes, G. E., & Cresswell, M. J. (1996). *A new introduction to modal logic*. Psychology Press.

Kshetri, N., & Voas, J. (2017). The economics of "fake news". *IT Professional*, 19(6), 8-12.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.

Lewis, D. (1973). Causation. *The Journal of Philosophy* (70:17), Seventieth Annual Meeting of the American Philosophical Association Eastern Division, 556-567.

Lewis, D. (1986). *Philosophical Papers*: Volume II. Oxford: Oxford University Press.

Marwala, T. (2014). *Artificial Intelligence Techniques for Rational Decision Making*. Springer, Heidelberg.

Marwala, T., and Hurwitz, E. (2017). Artificial Intelligence and Economic Theory: Skynet in the Market. DOI: 10.1007/978-3-319-66104-9_12

McKinnon, J. D., & Tracy, R. (2024, January 31). 'You Have Blood on Your Hands': Senators Say Tech Platforms Hurt Children. *Wall Street Journal*. https://www.wsj.com/tech/meta-tiktok-ceos-to-defend-against-claims-their-platforms-hurt-children-2c966c2b?mod=Searchresults_pos2&page=1

Mellor, D. H. (1995). *The Facts of Causation*, London: Routledge.

Mellor, D. H. (2004). For Facts as Causes and Effects. in Collins, J., Hall, N., and Paul, L. A. (eds.), pp. 309–24, *Causation and Counterfactuals*, The MIT Press. DOI: https://doi.org/10.7551/mitpress/1752.001.0001

Millican, P. (2007). David Hume: An Enquiry Concerning Human Understanding, Oxford University Press: Oxford.

Miranda, S. M., & Saunders, C. S. (2003). The social construction of meaning: An alternative perspective on information sharing. *Information systems research*, 14(1), 87-106.

Modal Logic. (Jan 23, 2023). Stanford Encyclopedia of Philosophy. Retrieved Dec 10, 2023, from https://plato.stanford.edu/entries/logic-modal/

Murphy, M. (2024, Jan 23). Biden Audio Deepfake Alarms Experts in Lead-Up to Elections. *Time*. https://time.com/6565446/biden-deepfake-audio/

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3-26.

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1), 1-39.

Ni, J., Rossiello, G., Gliozzo, A., & Florian, R. (2022). A Generative Model for Relation Extraction and Classification. arXiv preprint arXiv:2202.13229.

Paul, L. A., and Hall, N. (2013). Causation: A User's Guide. Oxford: Oxford University Press.

Pearl, J. (2000). *Causality*, Cambridge: Cambridge University Press.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11), 4944-4957.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Piaget, J. (1986). Essay on necessity. *Human development*, 29(6), 301-314.

Routley, R., & Routley, V. (1969). A fallacy of modality. *Nous*, 129-153.

Simon, H. A., American Economic Association, & Royal Economic Society. (1966). Theories of decision-making in economics and behavioural science (pp. 1-28). Palgrave Macmillan UK.

Stanford Encyclopedia of Philosophy. (2019). https://plato.stanford.edu/index.html

Starr, W. (2019). Counterfactuals. In Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/counterfactuals/

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, 25(2), 162-180.

Verma, P. (2023, December 17). The rise of AI fake news is creating a 'misinformation superspreader.' *The Washington Post*. https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Walter, N., Cohen, J., & Murphy, S. T. (2018). Self-affirmation activates the ventral striatum: A possible reward-related mechanism for self-affirmation. *Psychological Science*, 29(6), 1024-1033.

Walton, D. (1992). *Slippery slope arguments*. Oxford University Press.

Wang, S. A., Pang, M. S., & Pavlou, P. A. (2022). Seeing is believing? How including a video in fake news influences users' reporting of the fake news to social media platforms. *MIS Quarterly*, 46(3), 1323-1354.

Woodward, J. (1996). Explanation, Invariance, and Intervention. *Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association*. Part II: Symposia Papers, pp. S26-S41.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.

Woodward, J. and Hitchcock, C. (2003). Explanatory generalizations. Part I: a counterfactual account. *Noûs*, 37(1), 1–24.

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A unified generative framework for various NER subtasks. arXiv preprint arXiv:2106.01223.

Zhang, Y., Cao, D., & Liu, Y. (2022). Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. Advances in Neural Information Processing Systems, 35, 10643-10655.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5), 1-40.