MAFALDA Annotation Guidelines

December 14, 2023

TL;DR An argument consists of **one or more premises and a conclusion**. Premises and the conclusion can be **explicit or implicit**. **A fallacy is an invalid argument**, i.e., the premises do not entail the conclusion. False statements, generalizations, insults, slogans, and appeals to emotion are not fallacies unless used as premises in a fallacious argument. Fallacies are **not dependent on the truth of their components** but on the logical connection between premises and conclusion. The span of a fallacy in a text includes **all sentences with the conclusion and premises**. Conclusion and premises can be **pronouns**. Annotating fallacies in a dataset involves selecting relevant text spans and labeling them with appropriate fallacy types.

1 What is a fallacy?

Definition 1. *An argument says that one or more assertions called premises entail an assertion called the conclusion. Premises or conclusions can be implicit in an argument.*

Thus, an argument is always of the form "*A, B, C*, ... therefore *X*" or of the form "*X* because *A, B, C*, ...", or it can be rephrased into these forms. The premises or conclusion can also be implicit. Examples:

1. "***Premise***$_1$***:*** *All humans are mortal.* ***Premise***$_2$***:*** *Socrates is human.* ***Conclusion:*** *Therefore, Socrates is mortal.*"
2. "***Conclusion:*** *Socrates is mortal because* ***Premise***$_1$***:*** *he is a human and* ***Premise***$_2$***:*** *all humans are mortal*"
3. "*Of course,* ***Conclusion:*** *Socrates is mortal! How can you doubt this? After all,* ***Premise***$_1$***:*** *he's human!*" The second premise is implicit.

Definition 2. *A fallacy is an argument where the premises do not logically entail the conclusion. A fallacy belongs to a given category such as "Appeal to authority" (see Section 3).*

For example, *Einstein believed in God, and therefore God exists*, is a fallacy: From the fact that Einstein believed in God, it does not follow that God exists. It is an appeal to authority.**A fallacy is always an argument**. Thus, the following are **not fallacies**:

1. "*Paris is the capital of England*" is a false claim, but not a fallacy.
2. "*You are too stupid*" is an insult, but not a fallacy.
3. "*Think of the poor children!*" appeals to emotion, but is not per se a fallacy

A fallacy is **an argument that is invalid even if the premises were true**. Thus, the following are **not fallacies**:

1. "*During a Covid-19 pandemic, you should wear a mask in public transport because otherwise you could get infected*" is a valid argument because the premise does entail the conclusion.
2. "*All Americans love Trump, and therefore Biden loves Trump*" rests on a false premise (all Americans love Trump), but is not a fallacy. If the premise were true, then the conclusion would hold.

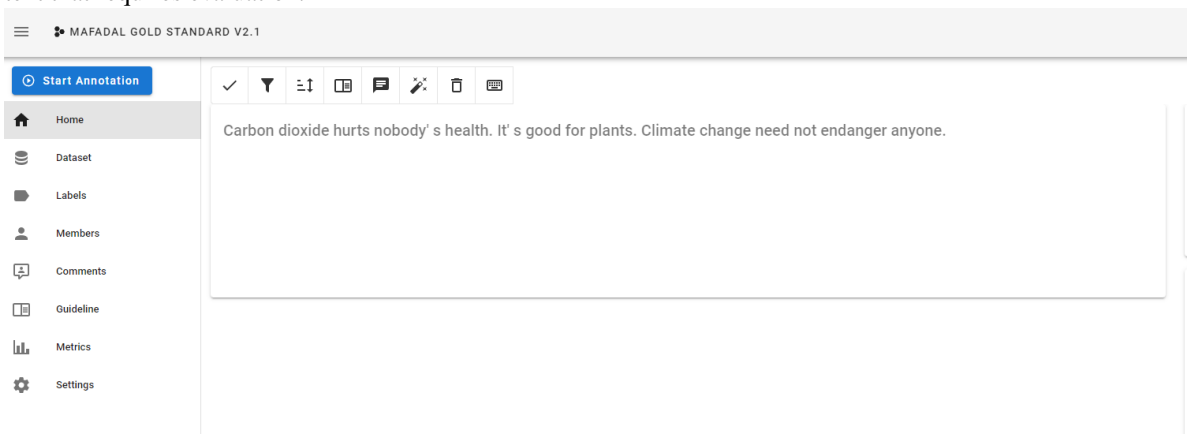
Definition 3 (Span)**.** *The span of a fallacy in a text comprises all sentences that include the conclusion and all given premises of the fallacy. A premise or conclusion need not be included if referred to by a pronoun. The conclusions or the premises can be implicit in the text, but they have to be reasonably understandable by a human.*

We work on the **level of entire sentences**, including punctuation. You don't have to annotate a sentence if you can annotate the pronoun instead.

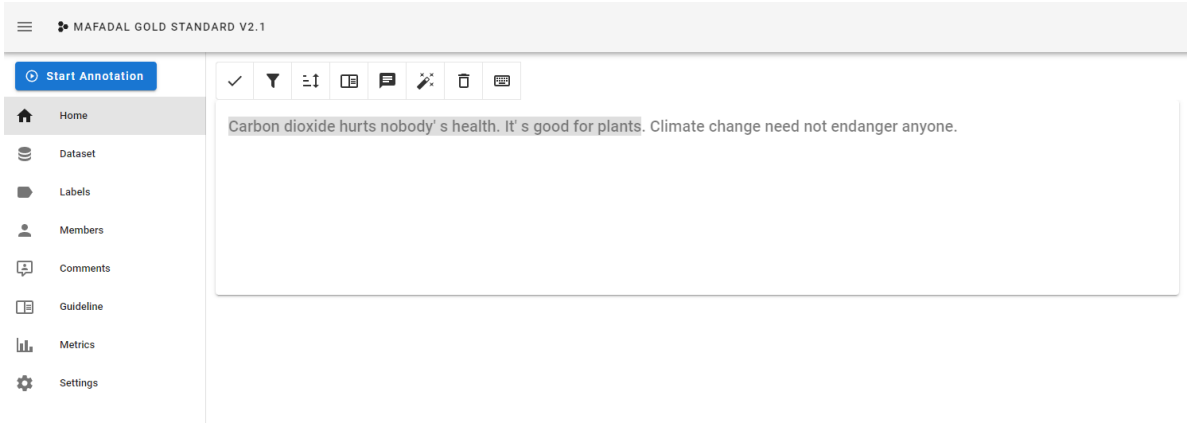
Example 1: Consider the following text: "*Can I get into finance with a Law degree? (...) This is law school arrogance at its finest. Why not a brain surgeon?*". This text is a fallacious argument: from the fact that one cannot become a brain surgeon with a law degree, the speaker concludes that one cannot work in finance with a law degree. The span of this fallacy is underlined: The conclusion is that "this" (working in finance with a law degree) is arrogant. The premise is that one cannot work as a brain surgeon with a law degree.

2 Instructions for Annotating the Dataset

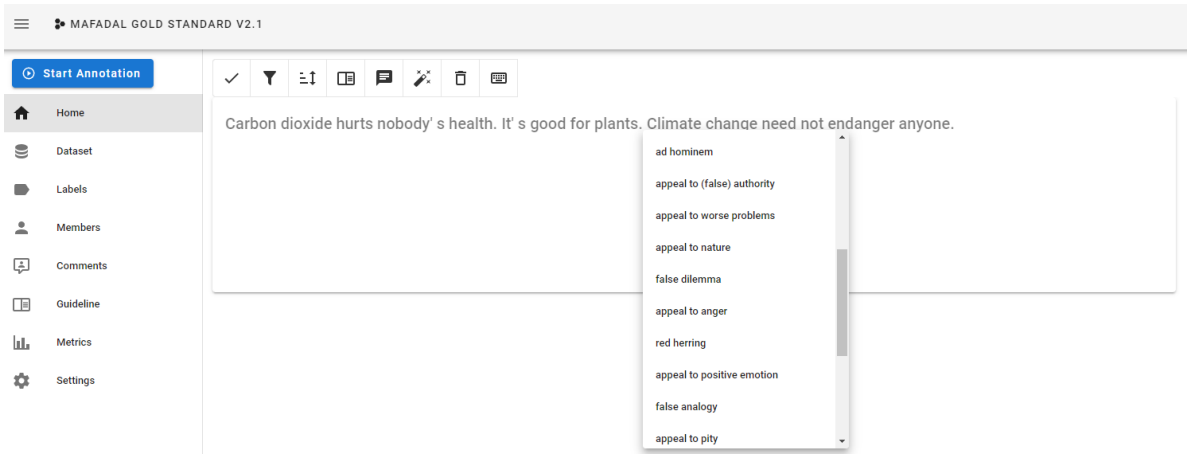
1. **Review the Example**
   Begin by examining the provided text in the dataset to understand the context. The example text will typically a text that requires evaluation.



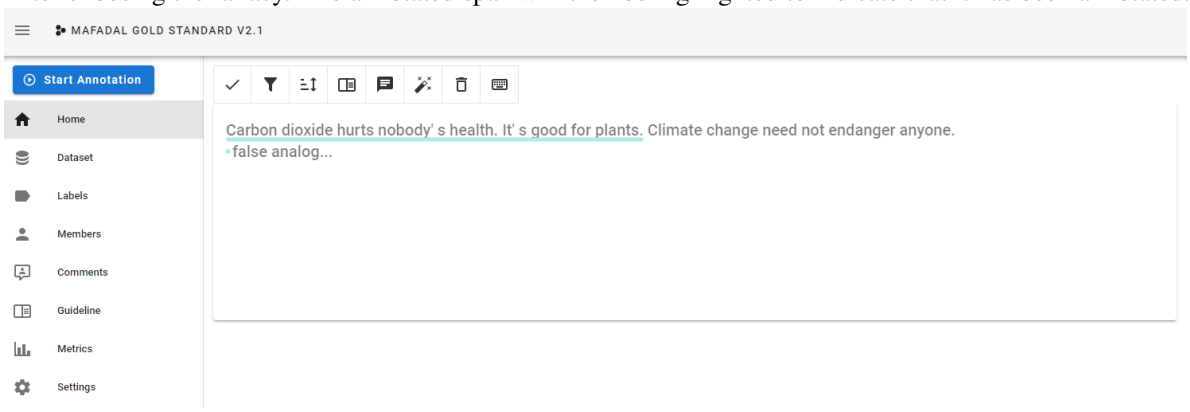
2. **Select Text Span**
   Click and drag to select the specific span of text within the example that you wish to annotate. A span can involve one or more sentences. Always mark entire sentences, including punctuation. Go to Section 1 to be sure that your example is a fallacy according to our definition



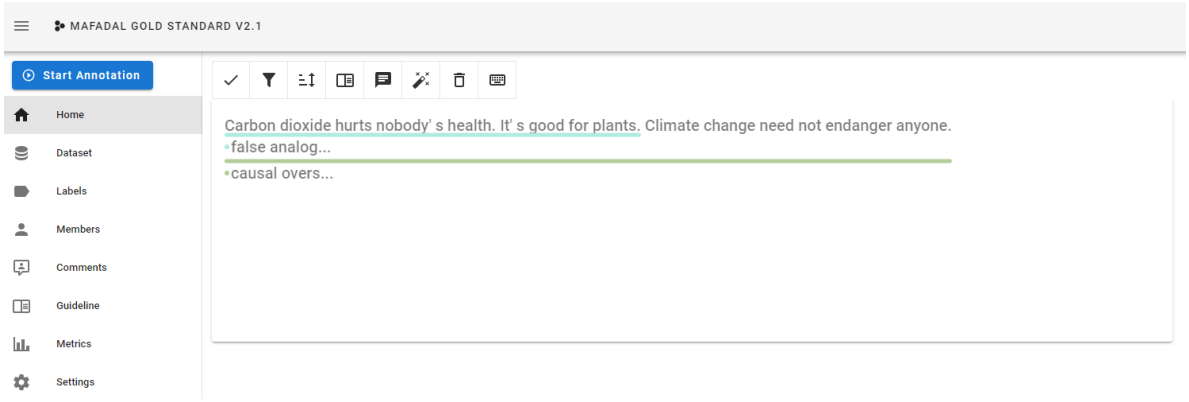
3. **Choose the Fallacy Category**
   Once a span is selected, a menu of fallacies will appear. Choose the one that best describes the fallacy with the selected text span. Go to Section 3 to be sure that you choose the correct fallacy.



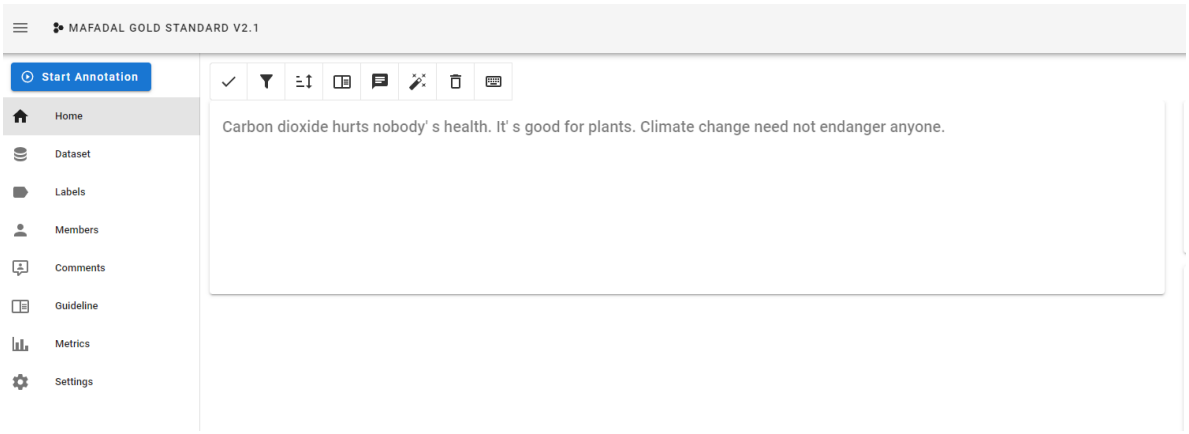
4. **Annotate Selected Span**
   After choosing the fallacy. The annotated span will then be highlighted to indicate that it has been annotated.



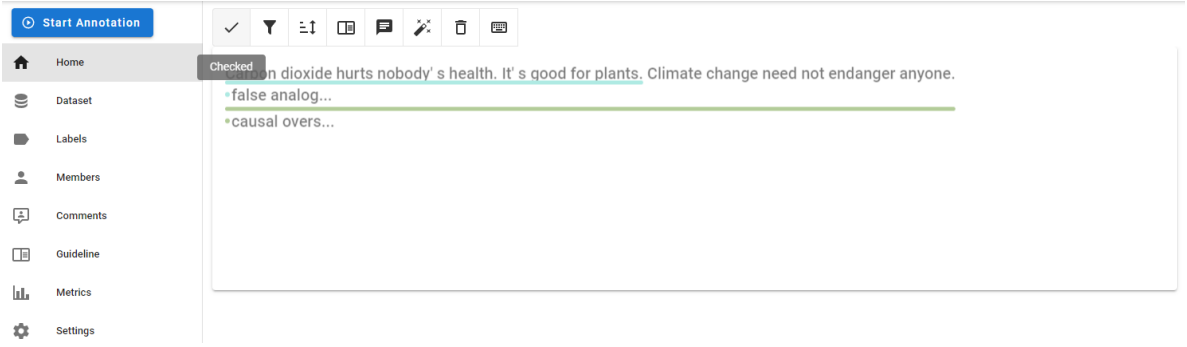
5. **Annotate Additional Spans**
   Continue to select and annotate additional text spans.



6. **Non Fallacious Text**
   If the text does not contain any fallacy, don't select any span and leave the text blank.



7. **Confirm Annotations**
   After annotating the necessary spans, confirm your annotations. This step involve clicking a *check* button to finalize the annotations. Ensure that all relevant text spans have been reviewed and appropriately annotated.



3 Definitions of the fallacies

In the following, we provide, for each fallacy, its informal definition, its formal definition, and a toy example. During the annotation you don't need to provide the annotation with variables.

We start by describing the variables/placeholders used in the formal templates.

- $A =$ attack
- $E =$ entity (persons, organizations) or group of entities
- $P, P_i =$ premises, properties, or possibilities
- $C =$ conclusion

Abusive Ad Hominem

Informal: This fallacy involves attacking a person's character or motives instead of addressing the substance of their argument.

Formal: E claims P. E's character is attacked (A). Therefore, $\neg P$.

Example: "John says the earth is round, but he's a convicted criminal, so he must be wrong."

Annotation with Variables: E (John) claims P (the earth is round). John's character is attacked (A) (being a criminal). Therefore, $\neg P$ (the earth is not round).

Ad Populum

Informal: This fallacy involves claiming that an idea or action is valid because it is popular or widely accepted.

Formal: A lot of people believe/do P. Therefore, P. OR Only a few people believe/do P. Therefore, $\neg P$.

Example: Millions of people believe in astrology, so it must be true.

Annotation with Variables: Many people believe in P (astrology). Therefore, P (astrology is true).

Appeal to Anger

Informal: This fallacy involves using anger or indignation as the main justification for an argument, rather than logical reasoning or evidence.

Formal: E claims P. E is outraged. Therefore, P. Or E_1 claims P. E_2 is outraged by P. Therefore, P (or $\neg P$ depending on the situation).

Example: The victim's family has been torn apart by this act of terror. Put yourselves in their terrible situation, you will see that he is guilty.

Annotation with Variables: E (the speaker) claims P (the accused is guilty) and expresses outrage. Therefore, P (guilt).

Appeal to Authority

Informal: This fallacy occurs when an argument relies on the opinion or endorsement of an authority figure who may not have relevant expertise or whose expertise is questionable. When applicable, a scientific consensus is not an appeal to authority.

Formal: E claims P (when E is seen as an authority on the facts relevant to P). Therefore, P.

Example: A famous actor says this health supplement works, so it must be effective.

Annotation with Variables: E (famous actor) claims P (the health supplement works). Therefore, P (it must be effective).

Appeal to Fear

Informal: This fallacy occurs when fear or threats are used as the main justification for an argument, rather than logical reasoning or evidence.

Formal: If $\neg P_1$, something terrible P_2 will happen. Therefore, P_1.

Example: If you don't support this politician, our country will be in ruins, so you must support them.

Annotation with Variables: If $\neg P_1$ (not supporting the politician), then P_2 (country in ruins) will happen. Therefore, P_1 (must support the politician).

Appeal to Nature

Informal: This fallacy occurs when something is assumed to be good or desirable simply because it is natural, while its unnatural counterpart is assumed to be bad or undesirable.

Formal: P_1 is natural. P_2 is not natural. Therefore, P_1 is better than P_2. OR P_1 is natural, therefore P_1 is good.

Example: Herbs are natural, so they are better than synthetic medicines.

Annotation with Variables: P_1 (herbs are natural) and P_2 (synthetic medicines are not natural), leading to P_1 is better than P_2.

Appeal to Pity

Informal: This fallacy involves using sympathy or compassion as the main justification for an argument, rather than logical reasoning or evidence.

Formal: P which is pitiful therefore C, with only a superficial link between P and C

Example: He's really struggling, so he should get the job despite lacking qualifications.

Annotation with Variables: P (he's struggling) is presented as a pitiful situation, leading to C (he should get the job), despite a superficial link between P and C.

Appeal to Positive Emotion

Informal: This fallacy occurs when a positive emotion – like hope, optimism, happiness, or pleasure – is used as the main justification for an argument, rather than logical reasoning or evidence.

Formal: P is positive. Therefore, P.

Example: Smoking a cigarette will make you look cool, you should try it!

Annotation with Variables: P (smoking cigarettes looks cool) leads to P (try smoking).

Appeal to Ridicule

Informal: This fallacy occurs when an opponent's argument is portrayed as absurd or ridiculous with the intention of discrediting it.

Formal: E_1 claims P. E_2 makes P look ridiculous, by misrepresenting P (P'). Therefore, $\neg P$.

Example: There's a proposal to reduce carbon emissions by 50% in the next decade. What's next? Are we all going to stop breathing to reduce CO2?

Annotation with Variables: E_1 (unspecified entity) claims P (proposal to reduce the carbon emissions). E_2 (the speaker) made P looks ridiculous by suggesting an extreme scenario P' (stop beathing). Therefore, $\neg P$ (reducing carbon emission is not reasonable).

Appeal to Tradition

Informal: This fallacy involves arguing that something should continue to be done a certain way because it has always been done that way, rather than evaluating its merits.

Formal: We have been doing P for generations. Therefore, we should keep doing P. OR Our ancestors thought P. Therefore, P.

Example: We've always had a meat dish at Thanksgiving, so we should not change it.

Annotation with Variables: P (always had meat dish at Thanksgiving) should continue. Therefore, continue P.

Appeal to Worse Problems

Informal: This fallacy involves dismissing an issue or problem by claiming that there are more important issues to deal with, instead of addressing the argument at hand. This fallacy is also known as the "relative privation" fallacy.

Formal: P_1 is presented. P_2 is presented as a best-case. Therefore, P_1 is not that good. OR P_1 is presented. P_2 is presented as a worst-case. Therefore, P_1 is very good.

Example: Why worry about littering when there are bigger problems like global warming?

Annotation with Variables: P_1 (littering) is compared to P_2 (global warming), which is a worse problem, leading to P_1 is not important.

Causal Oversimplification

Informal: This fallacy occurs when a complex issue is reduced to a single cause and effect, oversimplifying the actual relationships between events or factors.

Formal: P_1 caused C (although P_2, P_3, P_4, etc. also contributed to C.)

Example: There is an economic crisis in the country, the one to blame is the president.

Annotation with Variables: P_1 (the president) caused C (economic crisis), while ignoring other contributing factors (P_2 (worldwide economical context), P_3 (previous policies), etc.).

Circular Reasoning

Informal: This fallacy occurs when an argument assumes the very thing it is trying to prove, resulting in a circular and logically invalid argument.

Formal: C because of P. P because of C. OR C because C.

Example: The best smartphone is the iPhone because Apple creates the best products.

Annotation with Variables: C (IPhone is the best smartphone) because P (Apple creates the best products), which in turn is justified by the claim C.

Equivocation

Informal: This fallacy involves using ambiguous language or changing the meaning of a term within an argument, leading to confusion and false conclusions.

Formal: No logical form: P_1 uses a term T that has a meaning M_1. P_2 uses the term T with the meaning M_2 to mislead.

Example: The government admitted that many cases of credible UFOs (Unidentified flying objects) have been reported. Therefore that means that Aliens have already visited Earth.

Annotation with Variables: P_1 (many cases of credible UFOs have been reported) uses the term UFO with the meaning M_1 (unidentified flying objects). P_2 (aliens have already visited Earth) uses UFO with a different meaning M_2 (implying that aliens = UFOs), misleading the conclusion.

Fallacy of Division

Informal: This fallacy involves assuming that if something is true for a whole, it must also be true of all or some of its parts.

Formal: E_1 is part of E, E has property P. Therefore, E_1 has property P.

Example: The team is great, so every player on the team must be great.

Annotation with Variables: E_1 (every player) is part of E (the team). E has the property P (great), then E_1 also has P.

False Analogy

Informal: This fallacy involves making an analogy between two elements based on superficial resemblance.

Formal: E_1 is like E_2. E_2 has property P. Therefore, E_1 has property P. (but E_1 really is not too much like E_2)

Example: We should not invest in Space Exploration. It's like saying that a person in dept should pay for fancy vacations.

Annotation with Variables: E_1 (a state in debt plans to explore space) is linked to E_2 (a familly in dept plans fancy vacations). E_2 has property P (expensive and not advisable), implying E_1 should also have P.

False Causality

Informal: This fallacy involves incorrectly assuming that one event causes another, usually based on temporal order or correlation rather than a proven causal relationship.

Formal: P is associated with C (when the link is mostly temporal and not logical). Therefore, P causes C.

Example: After the rooster crows, the sun rises; therefore, the rooster causes the sunrise.

Annotation with Variables: P (rooster crows) is associated with C (sunrise), but the link is temporal, not causal, leading to the false conclusion that P causes C.

False Dilemma

Informal: This fallacy occurs when only two options are presented in an argument, even though more options may exist.

Formal: Either P_1 or P_2, while there are other possibilities. OR Either P_1, P_2, or P_3, while there are other possibilities.

Example: You're either with us, or against us.

Annotation with Variables: Presents a choice between P_1 (with us) and P_2 (against us), excluding other possibilities.

Guilt by Association

Informal: This fallacy involves discrediting an idea or person based on their association with another person, group, or idea that is viewed negatively.

Formal: E_1 claims P. Also E_2 claims P, and E_2's character is attacked (A). Therefore, $\neg P$. OR E_1 claims P. E_2's character is attacked (A) and is similar to E_1. Therefore $\neg P$.

Example: Alice believes in climate change, just like the discredited scientist Bob, so her belief must be false.

Annotation with Variables: E_1 (Alice) claims P (belief in climate change). E_2 (Bob) also claims P. However E_2's character (A) is attacked (being discredited). Therefore $\neg P$.

Hasty Generalization

Informal: This fallacy occurs when a conclusion is drawn based on insufficient or unrepresentative evidence.

Formal: Sample E_1 is taken from population E. (Sample E_1 is a very small part of population E.) Conclusion C is drawn from sample E_1.

Example: I met two aggressive dogs, so all dogs must be aggressive.

Annotation with Variables: A small sample E_1 (two aggressive dogs) is taken from a larger population E (all dogs). Therefore C (all dogs are aggressive).

Slippery Slope

Informal: This fallacy occurs when it is claimed that a small step will inevitably lead to a chain of events, resulting in a significant negative outcome.

Formal: P_1 implies P_2, then P_2 implies P_3,... then C which is negative. Therefore, $\neg P_1$.

Example: If we allow kids to play video games, they will see fights, guns, and violence, and then they'll become violent adults.

Annotation with Variables: P_1 (allowing kids to play video games) implies P_2 (seeing fights, guns, and violence), which in turns implies P_3 (to like violence, etc.) leading to C (kids becomes violent adults). Therefore, $\neg P_1$.

Strawman Fallacy

Informal: This fallacy involves misrepresenting an opponent's argument, making it easier to attack and discredit.

Formal: E_1 claims P. E_2 restates E_1's claim (in a distorted way P'). E_2 attacks (A) P'. Therefore, $\neg P$.

Example: He says we need better internet security, but I think his panic about hackers is overblown.

Annotation with Variables: E_1 (an unspecified person (He)) claims P (need for better internet security), E_2 (the speaker) distorts the claim as P' (panic about hackers). Therefore $\neg P$.

Tu Quoque

Informal: This fallacy occurs when someone's argument is dismissed because they are accused of acting inconsistently with their claim, rather than addressing the argument itself.

Formal: E claims P, but E is acting as if $\neg P$. Therefore $\neg P$.

Example: Laura advocates for healthy eating but was seen eating a burger, so her advice on diet is invalid.

Annotation with Variables: E (Laura) claims P (advocates for healthy eating), but E is acting as if $\neg P$ (eating a burger, which is unhealthy eating). Therefore $\neg P$ (advice on diet is invalid).