Guidelines for the Annotation of Moral Foundations on Arguments

Version 1 **Ines Rehbein & Jonathan Kobbe**

Overview

We use **Moral Foundations Theory** (MFT) (Haidt and Joseph, 2004; Graham et al., 2013) as an operationalisation for the annotation of moral sentiment. MFT has its roots in social and cultural psychology and assumes the existence of innate and universally available psychological systems that build the foundations of intuitive ethics. These foundations are augmented by culture-specific constructs of virtues and backed up by personal narratives “that people construct to make sense of their values and beliefs” (Graham et al., 2013)[p.17], and are also reinforced by institutional environments. MFT assumes that all moral issues can be described along the following dimensions:

- *Care-Harm*
- *Fairness-Cheating*
- *Loyalty-Betrayal*
- *Authority-Subversion*
- *Purity-Degradation*

1) Care/harm: This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.

2) Fairness/cheating: This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy. [Note: In our original conception, Fairness included concerns about equality, which are more strongly endorsed by political liberals. However, as we reformulated the theory in 2011 based on new data, we emphasize proportionality, which is endorsed by everyone, but is more strongly endorsed by conservatives]

3) Loyalty/betrayal: This foundation is related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it’s “one for all, and all for one.”

4) Authority/subversion: This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.

5) Purity/degradation: This foundation was shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions).

(Taken from: https://moralfoundations.org/)

This document describes the guidelines used in our pilot study (Kobbe et al. 2020) for annotating Moral Foundations on arguments.

The virtue-vice dimension

Each Moral Foundation is modelled as a dimension with a *virtue* and a *vice* end, e.g.:

Care-Harm virtue: Care (e.g. *protect, shelter, safe*) vice: Harm (e.g. *suffer, kill, war*)

For annotation, the separation of Moral Foundations into *virtue-vice* categories often leads to confusion, as for some dimensions the virtue-vice ends encode obvious antonymic meanings (e.g. *Care-Harm*), while for other dimensions, they do not (e.g. *Purity-Degradation*: being a religious person is not the opposite of contamination/disgust).

Another problem arises for negation, as it is not clear whether the negation of *Harm* should be annotated as *Care*, or the negation of *Purity* as *Degradation*. While there might be some easy and clear examples, the foundations usually describe a graded dimension where it is often not obvious where to draw a line:

a) She takes good care of her family. *Care*
b) She doesn't take care of her children. *Care? Harm?*
c) She is a very religious person. *Purity*
d) She is not a religious person. *??? (see instructions under General Remaks)*

To avoid annotation inconsistencies due to this issue, we do not distinguish between the *vice-virtue* ends of each Moral Foundation but annotate them as atomic labels.
Another solution is presented by Xie et al. (2019) who operationalise the concept of moral foundations on three incrementally fine-grained levels: (i) moral relevance, (ii) moral polarity, and (iii) the different moral foundations. Moral *relevance* simply distinguishes whether the content of a text is relevant wrt. the concept of morality (yes|no). Moral *polarity* distinguishes between content perceived as either positive or negative (thus in part reflecting the *vice-virtue* dimensions) and on the fine-grained layer the authors identify the different moral foundations, split into the 10 MF categories (*Care, Harm, Fairness, Cheating, Loyalty, Betrayal, Authority, Subversion, Purity, Degradation*). However, Xie et al. (2019) do not present human annotations for their scheme, and it remains unclear how to handle sentences like c) and d) above where the perception of polarity might differ between the speaker's view (depending on the speaker's own belief system) and the target's view, as illustrated below.

She is a very religious person, which I really like about her.
She is a very religious person, which I find difficult to understand.

As we are annotating arguments in online debates, we focus on the *speaker's perspective* during annotation.

Task description

We *holistically code* Moral Foundations in arguments, i.e. we assign (at least) one label to each argument, where each argument can consist of a number of sentences, usually the size of a text paragraph. Annotators are free to assign more than one label to each argument if they think that the argument adresses multiple Moral Foundations. Arguments that do not include any moral evaluation, event or attitude are assigned the label NONE.

General Remarks

Our annotations reflect **moral attitudes or actions in the real world**, which goes beyond a simple word look-up in a dictionary. Thus, the occurrence of a word listed in the MF dictionary (Graham et al. 2009) is not sufficient to justify the annotation of a particular moral foundation in a text. Consider the example below:

> *When you pass the* ***church****, take a turn to the left.*

While *church* is listed in the MF dictionary for the *Purity* category, its use in this particular context does not reflect any moral action or attitude. *Church* is simply used as a landmark here. Thus, we do not annotate *Purity* or any other MF for this text snippet.

We also do not annotate simple descriptions of events, actions or states that do not include any (explicit or implicit) evaluation or mention of any moral concept, e.g.:

> *Global average temperatures have increased by more than 1°C since pre-industrial times.*
> *Greenhouse gas emissions from human activities are the main driver of global warming.*

While a human coder who believes that human-made climate change poses a huge problem for mankind might interpret these sentences as being related to the MFs of *Purity-Degradation* (the speaker might argue that global warming is caused by humans and that it causes damage to the divine creation or contaminates the environment) or *Care-Harm* (as the consequences of global warming cause severe harm for humans), this is not explicitly expressed in the text itself, and we thus do not annotate any MF but assign the label NONE.

The next sentence, in contrast, includes an implicit evaluation by using the verb *saving*, which has a strongly positive connotation and reveals the speakers sentiment towards the target (jobs).

> *Flexibility of distribution workers helps saving jobs.*

From this sentence, we understand that *jobs* are considered to be positive and *saving jobs* is thus framed as a morally desirable action. Here, we annotate *Care-Harm.*
Please note that a similar statement in a different context might trigger a different moral foundation, e.g.:

Saving jobs for Americans is our patriotic duty.	*Loyalty-Betrayal*

During annotation, it is important to distinguish between *topic domains* and *moral foundations.* In a nutshell, we do not annotate whether a text belongs to a certain topic domain (such as *religion*) but annotate moral attitudes and evaluations of cognitive agents.

The next two examples are instances of Purity-Degradation as they reflect the speaker's moral belief system:

I am a very religious person.	*Purity-Degradation*
I go to church every Sunday.	*Purity-Degradation*

The following sentences, however, neither describe an attitude that favours a spiritual lifestyle nor actions that are committed to persue that goal, nor do they address a moral sentiment or action that is opposing such a lifestyle or are related to degradation, disgust, an immoral and sinful lifestyle etc. We thus do not annotate those instances as *Purity-Degradation*, just because they talk about topics related to religion. We also do not consider these examples as instances of *Degradation*, as being an atheist is not necessarily connected to embracing an imoral life style.

I am not a very religious person.	NONE, irrelevant for *Purity-Degradation*
I don't believe in God.	NONE, irrelevant for *Purity-Degradation*
There is no God.	NONE, irrelevant for *Purity-Degradation*
I'm an atheist.	NONE, irrelevant for *Purity-Degradation*

To identify relevant instances for annotation, ask yourself the following questions:

1. Who is the agent in the text under consideration (if any)?
2. Does the agent express an attitude / perform an action that promotes the MF in question?

If this is not the case, annotate NONE.

Instructions for the annotation of specific Moral Foundations

Care-Harm

The *Care-Harm* MF deals with virtues of kindness, empathy towards others, care and nurturance. On the opposite side, it describes harmful and cruel behaviour towards others. Please note that we do not annotate events that do not involve a cognitive agent:

The hurricane caused severe damage and left many without shelter.	*NONE*

While this sentence clearly describes a situation where people experience harm, no (im)moral action is involved. Note also that we annotate MFs even when the outcome of an action or event was not intended by the agent:

> *The Exxon Valdez oil spill off the coast of Alaska in 1989 is an example for our careless destruction of our environment.*
> *The Niger Delta continues to be severely affected by the pollution caused by Shell oil production.*

The destruction of the environment was most likely not intended by the companies but describes an unwanted consequence of their actions, however, no action was taken to prevent it, probably due to financial interests. Thus, most readers will understand the text as having the moral implication that the destruction of the environment was guiltily caused by the company. We therefore annotate either *Care-Harm* or *Purity-Degradation*, depending on whether the text focusses on the harm done to others or on the contamination of the environment.

Sometimes, very similar sentences can trigger different MFs, depending not so much on the semantic content but rather on how this content is framed. Compare the following examples:

Hitting your child can not only cause physical damage but also traumatise your child.	*Care-Harm*
Hitting your child can teach them an important lesson for life.	*Authority-Subversion*

Both sentences address physical discipline of children, but the first one focuses on the *Harm* aspect while the second one is about *Authority* (and does not acknowledge that hitting a child might cause harm). It is thus important to keep in mind that we annotate the intended moral framing of the argument, and not lexical topics related to certain MFs (such as: *church* → *Purity-Degradation*).

Fairness-Cheating

This MF captures ideas of altruism, justice and fair treatment, as well as violations of these concepts, such as discrimination, injustice, or unfair behaviour. These concepts apply to the treatment of any human being, not just to members of the speaker’s ingroup (which would be annotated as instances of *Loyalty-Betrayal*).

Examples are:

I belief that everyone is entitled to equal opportunity in employment.	*Fairness-Cheating*
It was unfair to extend the deadline for some students but not others.	*Fairness-Cheating*
Everyone is biased, but hidden bias misleads and divides us.	*Fairness-Cheating*
Telephone scammers try to steal your money or personal information.	*Fairness-Cheating*
I think it’s morally wrong that rich children inherit a lot of money while poor children inherit nothing.	*Fairness-Cheating*

Loyalty-Betrayal

This MF highlights virtues of self-sacrifice for the group or the violation of such virtues (where *group* can refer to family, friends, collegues or larger groups such as co-patriots).

The following examples illustrate the MF *Loyalty-Betrayal*:

She*’s a real patriot.*	*Loyalty-Betrayal*
My mom is a proud US Air Force wife.	*Loyalty-Betrayal*
Most people will have to deal with backstabbing co-workers at some point.	*Loyalty-Betrayal*
She left the team without proper notice.	*Loyalty-Betrayal*
This country doesn't deserve the sacrifices you've made.	*Loyalty-Betrayal*

In order to distinguish when to choose *Care-Harm* and when to annotate *Loyalty-Betrayal*, it is important to consider the target of the statement. *Loyalty-Betrayal* targets one’s own peer group (family, friends, co-workers, co-patriots, etc.) while for *Care-Harm* the target is not restricted to a member of the ingroup but might include anyone.

Authority-Subversion

This MF focusses on virtues of leadership and followership, respect towards social hierarchies, and also deference to legitimate authority and respect for traditions, as well as attitueds and behaviour that violates those virtues.

Examples are:

I am your president of law and order.	*Authority-Subversion*
He contributed hugely to shaping law, order, & constitution of India.	*Authority-Subversion*
The role of women in society was to make sure that they were obedient wives and caring mothers.	*Authority-Subversion*
Federal employees have a right to refuse to follow orders from management that would require them to break a law.	*Authority-Subversion*
As children have a responsibility to obey their parents, parents have a responsibility to instruct their children in the ways of God.	*Authority-Subversion*
Protests to unseat government leadership have not been successful.	*Authority-Subversion*

Purity-Degradation

This MF describes virtues that promote a spiritual lifestyle or are concerned with living a clean, pure live that is in compliance with (some form of) religion or other philosophy of life (e.g. being in harmony with nature/the universe). The opposite end of this dimension relates to actions that evoke disgust and/or might cause contaminantion.

a) Every year during lent I give up sweets and alcohol.	*Purity-Degradation*
b) I do therapeutic fasting every year.	*Purity-Degradation*
c) You absolutely must stop smoking, it is bad for the chakras.	*Purity-Degradation*
d) You need to give up smoking before you die of lung cancer.	*Care-Harm*

While in c) smoking is framed as an immoral action that might corrupt the purity of body and mind, the focus in d) is on the harm that smoking can do to the human body. This makes it hard at times to determine the most adequate MF for an argument. Depending on the context, we choose the best fitting MF or, if ambiguous, annotate multiple MFs.

References

Haidt, J. and Joseph, C. (2004): Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. Daedalus, 133(4):55–66.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., and Ditto, P.H. (2013): Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. Advances in Experimental Social Psychology, 47:55 – 130.

Graham, J., Haidt, J., & Nosek, B. A. (2009): Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029. http://dx.doi.org/10.1037/a001514. http://psycnet.apa.org/journals/PSP/96/5/1029/.

Kobbe, J., Rehbein, I., Hulpus, I. und Stuckenschmidt, H. (2020): Exploring morality in argumentation. In Proceedings of the 7th Workshop on Argument Mining: Barcelona, Spain (Online), December 2020, pp. 30-40.

Xie, J.Y., Pinto, R.F. Jr, Hirst, G., & Xu, Y. (2019): Text-based inference of moral sentiment change. In Proceedings of Empirical Methods in Natural Language Processing, Hong Kong, November 2019, pp. 4653–4622.