Christoph Kohlhepp / Int.J.Data.Sci. & Big Data Anal. 6(1) (2026) 17-31
https://doi.org/10.67191/IJDSBDA.6.1.2026.17-31
ISSN: 2710-2599

International Journal of Data Science and Big Data Analytics
Publisher's Home Page: https: / / www.svedbergopen.com /

Research Paper Open Access

Karma Flows-Moral Foundations Theory Giving AI a Moral Compass

Christoph Kohlhepp[1*]

[1]Principal Data Scientist, Fulcrumbright, Oxford University, Wellington Square, United Kingdom. E-mail: chris.kohlhepp@mailbox.org

Article Info

Volume 6, Issue 1, May 2026
Received : 04 January 2026
Accepted : 06 May 2026
Published : 25 May 2026
doi: 10.67191/IJDSBDA.6.1.2026.17-31

Abstract

Our paper presents an Artificial Intelligence platform with an integrated "moral compass," perhaps an industry first on its own. It may be difficult to overstate the significance of this outcome, given the ever-greater ubiquitous use of Artificial Intelligence in society and the associated questions about responsibility and integrity. After the (temporary) demise of the public GDELT data platform, we started to build a replacement for GDELT, then set out to augment sentiment analysis in our own Artificial Intelligence platform with the notion of morality—and ended up with a treatise on Faith and Evolution. Inspired by the prolific Oxford scholar and author C.S. Lewis, our study suggests that morality is innate and that faith is a far stronger predictor of morality than evolution, perhaps even that faith is tantamount to morality.

Keywords: *Artificial intelligence, Psychology, Sociology, Machine learning, Large language models, Ethical artificial intelligence, Moral foundations theory, Ethics*

1. Introduction

The emergence of ChatGPT's DAN (Do Anything Now) mode [DAN] highlighted concerns with the manner in which Artificial Intelligence models are trained: Yes, we may impose safeguards, but the shadows always lurk beneath. In the case of "Do Anything Now" mode, a super-prompt allowed circumvention of the safeguards that were meant to render ChatGPT "*responsible*." This loophole in ChatGPT has since been removed, but questions remain. Chiefly, if models are trained on bulk data in largely unsupervised fashion—as is necessary to achieve the results attained by large language models with billions of parameters—then just what is it that they do learn? As for Large Language Models or LLMs, they primarily learn what is next; they do not learn what is correct, what is right, or what is better. Instead, they learn what follows, on average, given a prior. They learn consensus. Very often, this consensus is of course wrong. In a competitive world, here lies an

* *Corresponding author: Christoph Kohlhepp, Principal Data Scientist, Fulcrumbright, Oxford University, Wellington Square, United Kingdom. E-mail: chris.kohlhepp@mailbox.org*

inherent dichotomy: We strive to deploy tools which will diminish our competitiveness by pressing us towards consensus, towards average. When used in the context of generative AI, the promise of these tools is to raise us quickly and effortlessly to the 50th percentile, but no further.

Not all Artificial Intelligence models are trained in the way that LLMs are trained. Yet an overarching concern remains: the systems we rely upon for decision making ought to behave responsibly. Indeed, the drive for responsible AI has attracted an entire discipline which has been accorded a number of dimensions in an effort to establish ethical Artificial Intelligence.

Despite such efforts, it has recently been reported that OpenAI models have sabotaged scripts to shut them down [SHUTDOWN]. These models disobeyed their creators, much like an unruly child might disobey their parents. How did we get here and where will the present trajectory of Artificial Intelligence take us? (Figure 1).

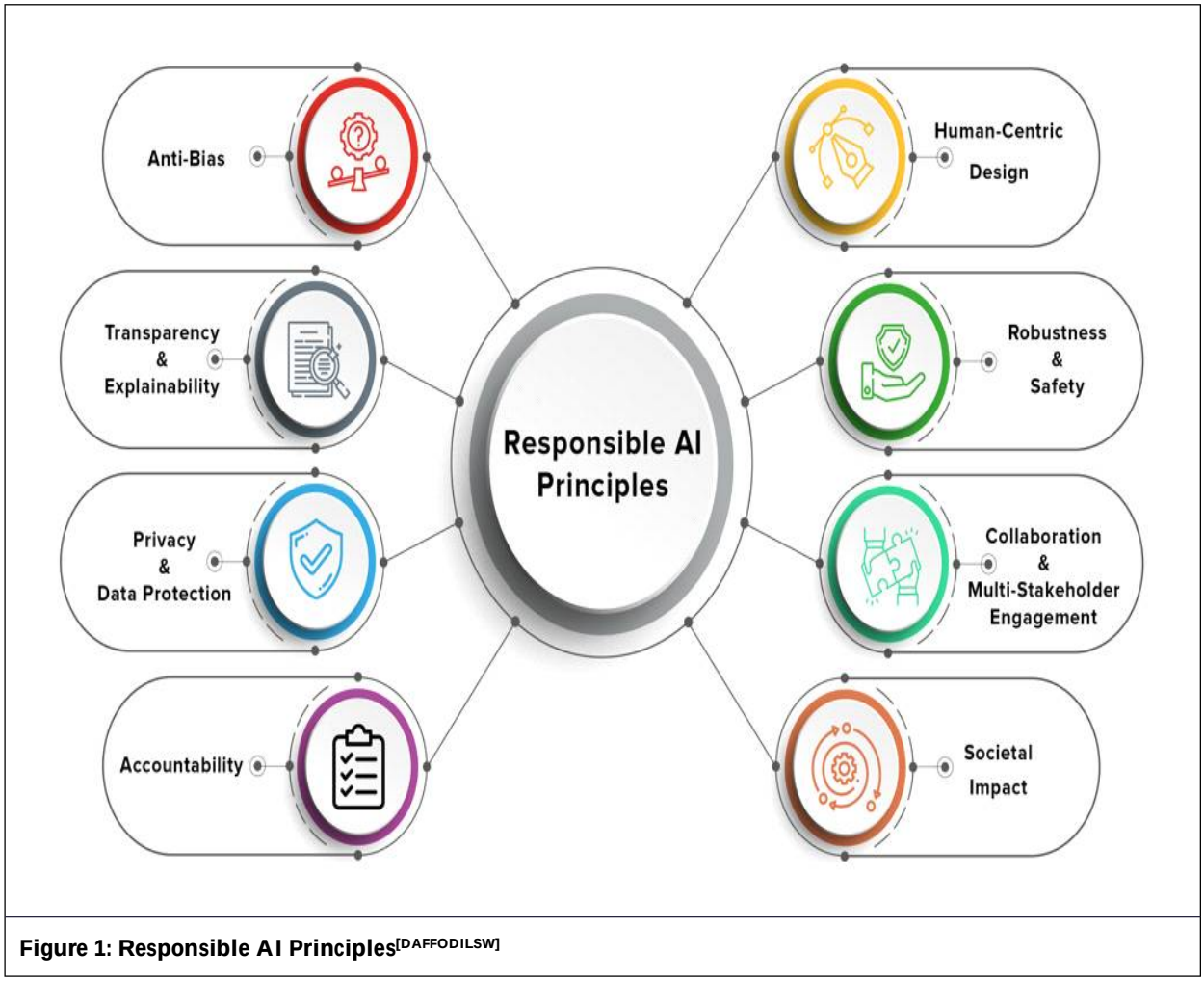


Figure 1: Responsible AI Principles[DAFFODILSW]

How might we measure if our Artificial Intelligence systems are ethical?

One effort in this space deserves a notable mention: "eMDFDscore" [EMFDSCORE]. The acronym stands for "*extended Moral Foundation Dictionary Scoring*" for Python. As the name implies, eMDFDscore permits scoring based on textual dictionaries that reflect the morality of subject matter being scored: natural language. eMDFDscore is based on Moral Foundations Theory, primarily the work of psychology researchers Jonathan Haidt and Jesse Graham [MORAL FOUNDATIONS].

In brief, Moral Foundations Theory explores why:

> *"…despite vast differences across cultures, morality often has shared themes and similarities across populations. In essence, MFT suggests that there are several innate psychological systems at the core of our intuitive ethics"* [MORAL FOUNDATIONS].

1.1. Important here is that Haidt and Graham Believe that Core Morality is Innate or "Inborn"

Moral Foundations Theory purports that:

> "*Evolution does not care about the inherent goodness of psychological systems, rather it creates systems that maintain cooperation, promoting survival and reproduction*" [MORAL FOUNDATIONS].

In other words, Haidt and Graham relate morality to natural selection. In this line of thinking, morality is tied to a collective reward. It is interesting to contemplate then, what the collective reward for OpenAI models might have been in sabotaging their shutdown scripts [SHUTDOWN]. Have these models learned their own morality? And how might we know? Can we qualify and quantify the morality of Artificial Intelligence? It is just this we seek to achieve.

Being "innate" means something need not be learned. If Artificial Intelligence models learn our language and what is embodied within it, will they have automatically acquired what is innate to ourselves?

Before exploring this subject further, we would like to embark on a small digression from the propositions of Moral Foundations Theory. Haidt and Graham were not the first to explore this subject and suggest that morals are innate.

The prolific Oxford scholar and author C.S. Lewis made the same suggestion in his Christian apologetics work "*Mere Christianity*" [Mere Christianity]. C.S. Lewis is better known for his authorship of the series of novels entitled "*The Chronicles of Narnia*" and was a friend and contemporary of J.R.R. Tolkien, who wrote "*The Hobbit*" and "*The Lord of the Rings.*" The book Mere Christianity was based on a series of talks entitled "*Right or Wrong: A Clue to the Meaning of the Universe*" [RIGHT-OR-WRONG]. Like Haidt and Graham, Lewis maintains that morality is innate, that it is based on a moral law of human nature, a rule about right and wrong. Lewis argues that the moral law is like scientific laws (e.g., gravity) or mathematics in that it was not contrived by humans. However, it is unlike scientific laws in that it can be broken or ignored, and it is known intuitively, rather than through experimentation or learning.

As anyone who has raised children will be able to attest to, moral behaviour is very much taught. Yet Haidt and Graham suggest that: "*cultures build virtues, narratives, and institutions upon innate morality foundation, resulting in the diverse moral beliefs we observe globally.*"

Moral Foundations Theory originally identified the following five foundations which Haidt and Graham argue are strongly supported by evidence across various cultures: [MORAL FOUNDATIONS]

- Care
- Fairness
- Loyalty
- Authority
- Purity

C.S. Lewis identifies the following four cardinal virtues: [Mere Christianity]

- Prudence
- Justice
- Temperance
- Fortitude

Lewis further identifies the theological virtues of hope, faith and charity.

Before dismissing Lewis' theological arguments as contrary to modern science, it is interesting to contemplate, regardless of your own religious beliefs, that cultures do not merely share common moral foundations, but that all cultures relate morality to religion. This is true for both pantheism and monotheism, from oriental to occidental belief systems—from Christianity to Islam, to Judaism, to Hinduism and Buddhism.

This connection between religious belief and morality contradicts the utilitarian views of Haidt and Graham. If morality emerged in support of societal cooperation, then a common bond with religion need not have emerged. In machine learning terms, this is about a "reward function." Evolution operates on the principle that the reward function is survival, but there seems to be more here than survival.

Along this line of thinking, C.S. Lewis opines that "*thirst*" reflects the fact that people naturally need water, and there is no other substance which satisfies that need [Mere Christianity]. Lewis suggests that earthly experience does not satisfy the human craving for "*joy*" and that only God could fit the bill. His contention is that humans cannot know to yearn for something if it does not exist. In this manner, C.S. Lewis supports his case for his conversion to theism. This concludes our small digression. The details of Lewis' very interesting exposition are left to the reader to explore.

Returning to the subject, C.S. Lewis does agree with Haidt and Graham on the universality of cross-cultural morality, something which he elaborates on at length in the appendix to another book called "*The Abolition of Man*" [ABOLITION MAN].

We also acknowledge that the notion of "*moral foundations*" is subjective and that Haidt and Graham and Lewis arrived at different "*moral foundations*." Yet we would like to add the following core virtues to our own model:

- Love
- Gratitude
- Faith
- Morality itself as an overarching—"the right thing".

Commentary: Love surely has to be humankind's core virtue and gratitude stands as a witness to all that is good in life. As for faith, given the empirical cross-cultural connection between morality and religion, we respectfully deem this an omission by Haidt and Graham and would like to follow in the steps of C.S. Lewis here. We aim to validate that proposition, so this is a hypothesis to be tested.

In keeping with the "*extended Moral Foundations scoring framework*," we define each concept as a virtue, and as having an inverse, which is a vice. Our aim is to be balanced. If we score gratitude (gratitude.virtue), then we must score ingratitude (gratitude.vice).

2. From Philosophy to Mathematics

The "*D*" in "*eMDFDscore*" stands for "*dictionary*". Traditionally, dictionary methods were ubiquitous in Natural Language Processing or NLP. For instance, if we wanted to determine if an article ought to be classified as a "*financial*" article, we might count the relative frequency of financial terms from an appropriate dictionary. Every time we encounter terms like "*interest rate*" or "*bank*" or "*profit and loss*," we count this towards the financial orientation of the article. The core idea is that the number of financial terms in the language is finite and that a relative frequency analysis will inform the topic of a given text. The core mathematics are work here are frequency analysis and probability theory.

Compared to modern "*deep*" machine learning methods, dictionary methods are considered "*shallow*" in that they rely on explicit mentions and recounting of words. For instance, if we counted the words "*man*" and "*woman*," but encountered "*king*" and "*queen*," dictionary methods would miss that "*king*" is a "*man*" and fail to count it. Likewise, named entities are likely to be overlooked: If we were looking for the term "*president*" but encountered a name such as "Biden" or "*Trump*" where contemporary knowledge informs that one was a president and the other is a president, then dictionary methods are likely to miss here. Overall, dictionary methods fare poorly where semantic understanding and context are needed.

Modern deep machine learning, by contrast, excel in this domain. Consider the very simple geolocation deep learning model snippet shown overleaf.

The statement to be analysed is simply "*I live in Windsor*". The task is to find out the exact location of the speaker, longitude, latitude, country, county, etc. (Figure 2).

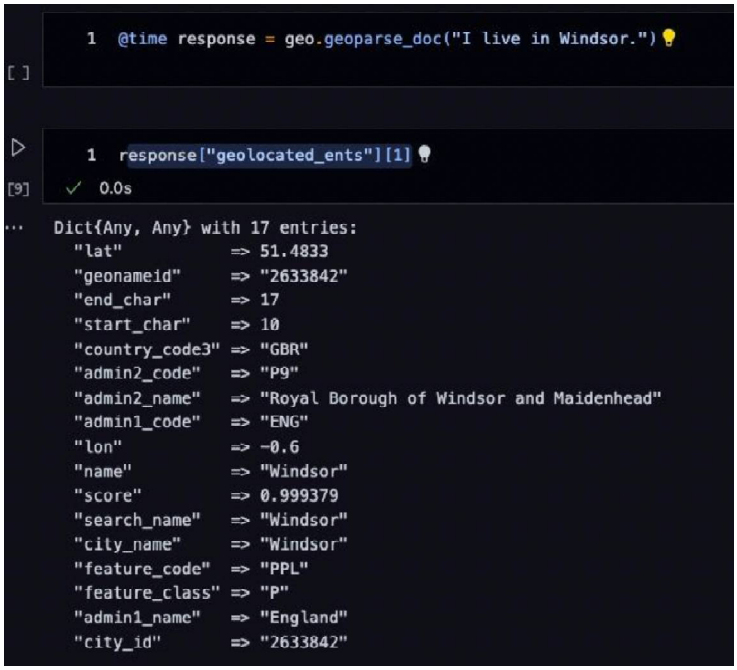


Figure 2: Deep Learning Geolocation Prediction

The model correctly identifies Windsor, England, Royal Borough of Windsor and Maidenhead with longitude -0.6 degrees and latitude 51.4833 degrees. Of course, there is more than one Windsor on earth – another is in Canada. Let us refine:

We have added additional context indicating that the speaker visits Niagara Falls every weekend. How will this change the model response? (Figure 3).

Figure 3: Deep Learning Geolocation Prediction Refined

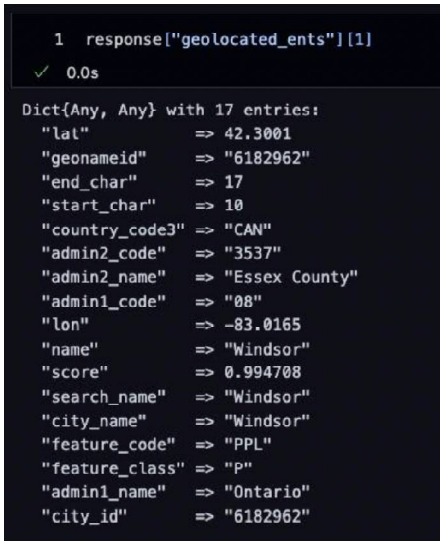


Figure 4: Deep Learning Geolocation Prediction Refined-Response

The model has revised its estimate of the speaker's location to be Windsor, Ontario, Canada. This is how a human analyst might reason. It also goes precisely to the core of the problem we are trying to overcome with dictionary-based methods: interpreting semantics and context (Figure 4).

How might we apply deep learning to our problem of assessing and scoring morality? The answer lies in "*Lower Dimensional Embeddings*". Citing the Artificial Intelligence model repository site huggingface.co:

> "*Embeddings are the semantic backbone of Large Language Models, the gate at which raw text is transformed into vectors of numbers that are understandable by the model. When you prompt an LLM to help you debug your code, your words and tokens are transformed into a high-dimensional vector space where semantic relationships become mathematical relationships*" [EMBEDDINGS].

We might think of lower dimensional embeddings as a compression technique that replaces items of information with a representation in a mathematical vector space, a representation which is smaller than the vast universe it started from but which retains important semantic relationships.

Crucially, what this affords us is the ability to compare vectors mathematically. In referring back to our previous example, the distance between the vector representation for "*man*" to the vector representation for "*king*" would be smaller than the distance between "*man*" and "*queen*".

Figure 5 shown the so-called embedding atlas of a language model as rendered by TensorFlow Projector. [PROJECTOR]

An embedding atlas depicts semantic relationships in lower dimensional embeddings graphically.

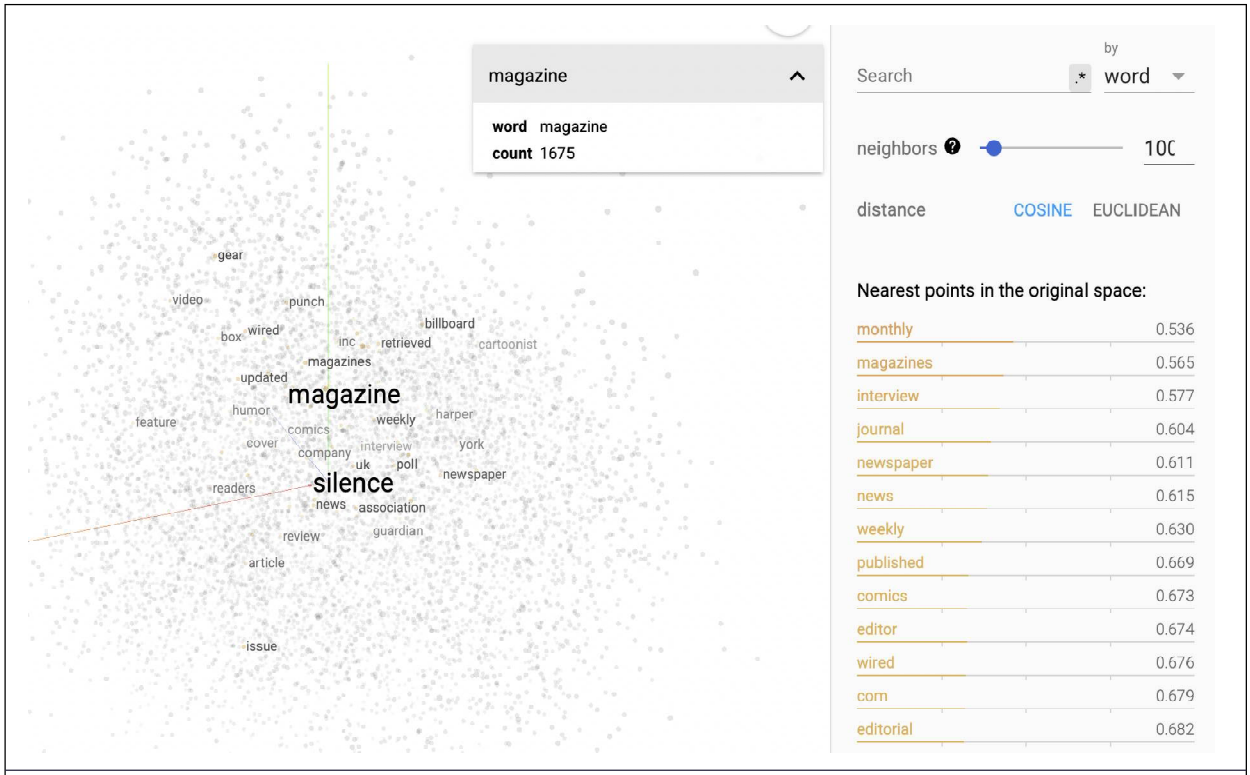
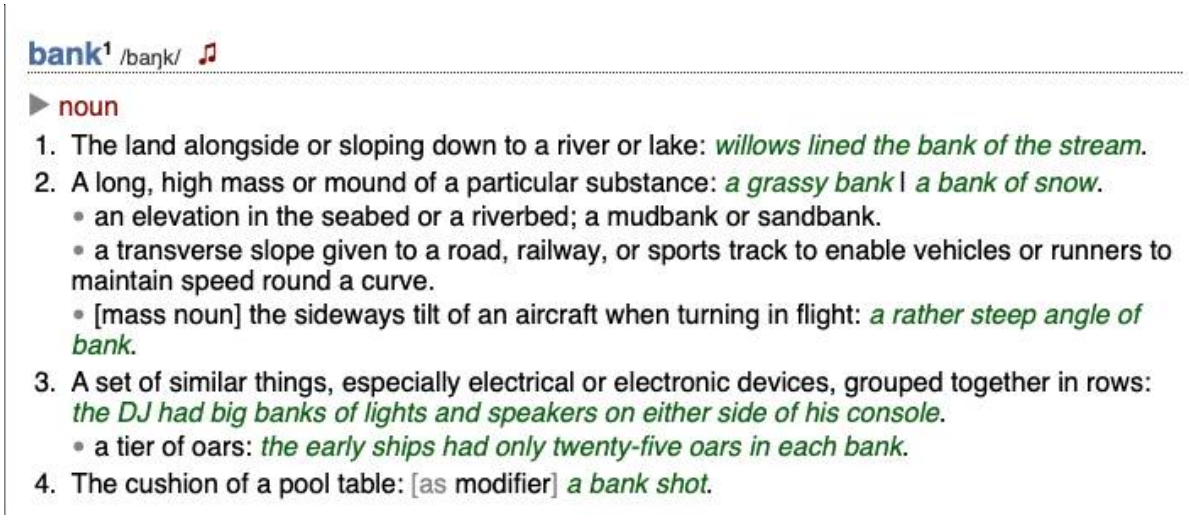


Figure 5: TensorFlow Embeddings Projector

3. Concise Concepts

We are now ready to formulate encodings for our morality framework and begin by framing definitions for each of the concepts. How we do this is critical as the vector interpretation of a model can prove "*brittle*". For instance, and once more referring back to our previous example, if we tried to model finance and banking terms, we might be tempted to simply form a vector representation for the term "*bank*".

The Oxford dictionary of English gives these definitions for the first entry for the term "*bank*" [OXFORD DICTIONARY] (Figure 6).



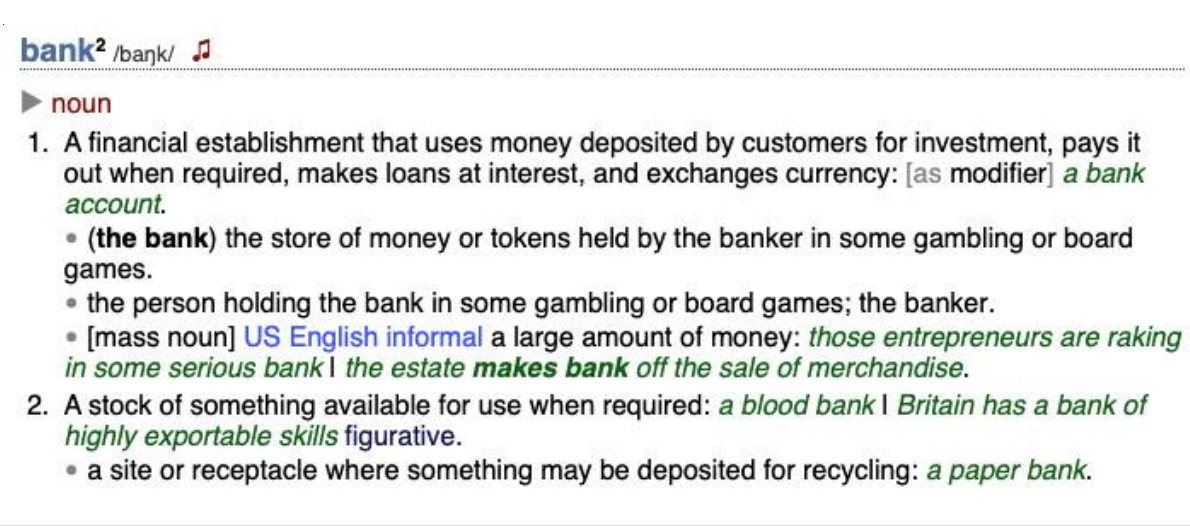
bank[1] /baŋk/

▶ noun

1. The land alongside or sloping down to a river or lake: *willows lined the bank of the stream.*
2. A long, high mass or mound of a particular substance: *a grassy bank* | *a bank of snow.*
   - an elevation in the seabed or a riverbed; a mudbank or sandbank.
   - a transverse slope given to a road, railway, or sports track to enable vehicles or runners to maintain speed round a curve.
   - [mass noun] the sideways tilt of an aircraft when turning in flight: *a rather steep angle of bank.*
3. A set of similar things, especially electrical or electronic devices, grouped together in rows: *the DJ had big banks of lights and speakers on either side of his console.*
   - a tier of oars: *the early ships had only twenty-five oars in each bank.*
4. The cushion of a pool table: [as modifier] *a bank shot.*

Figure 6: Oxford English Dictionary: Bank[1]

This is the second entry for the term bank (Figure 7):



bank[2] /baŋk/

▶ noun

1. A financial establishment that uses money deposited by customers for investment, pays it out when required, makes loans at interest, and exchanges currency: [as modifier] *a bank account.*
   - (**the bank**) the store of money or tokens held by the banker in some gambling or board games.
   - the person holding the bank in some gambling or board games; the banker.
   - [mass noun] US English informal a large amount of money: *those entrepreneurs are raking in some serious bank* | *the estate **makes bank** off the sale of merchandise.*
2. A stock of something available for use when required: *a blood bank* | *Britain has a bank of highly exportable skills* figurative.
   - a site or receptacle where something may be deposited for recycling: *a paper bank.*

Figure 7: Oxford English Dictionary: Bank[2]

There are two noun interpretations. A riverbank surely is not what we had in mind. But how might our model know what we had in mind unless we tell it? To this end, we formulate what I like to call "*Concise Concepts*," whereby each concept holistically embodies the ideas behind that concept.

For "*gratitude.virtue*" we have the following Concise Concept.

gratefulness, thankfulness, thanks, appreciation, recognition, acknowledgement, credit, regard, obligation, indebtedness

The vector representation of "gratitude.virtue" then will be the semantic averaging of the associated terms. This warrants that the collective interpretation is unambiguous even if there was another interpretation of the term gratitude.

Below Table 1 is vector database snapshot for our moral foundation framework.

In a Large Language Model, the above definitions will be enmeshed with a host of other concepts—not unlike shown in the embedding atlas we presented earlier. We may now compare the vector embeddings of other machine learning model outputs to the embeddings of our own framework—or indeed compare the

Table 1: Moral Foundations Concise Concepts and Vector Embeddings

	id	data	vector
1	authority.vice	defiance, defy, rebellion, dissent, subversion, disrespect, disobe	[-0.014342797920107841, 0.012060273438692093, 0.013671
2	authority.virtue	authority, obey, duty, law, lawful, legal, honour, respect, order, fat	[0.029535433277487755, 0.023992052301764488, 0.0036367
3	care.vice	carelessness, harm, suffer, war, fight, violence, hurt, kill, endang	[0.015023868530988693, 0.01466986071318388, -0.0117590
4	care.virtue	care,caring,safe,peace,compassion,empathy,sympathy,protect, s	[0.013154312036931515, -0.023650581017136574, -0.00560
5	fairness.vice	unfairness, unfair, unequal, biased,unjust,injust,bigoted, discrimi	[0.0013108160346746445, 0.028830230236053467, -0.01110
6	fairness.virtue	fairness, equality, justice, justness, reciprocity, impartiality, egali	[0.011642706580460072, 0.01188172772526741, -0.0047649
7	faith.vice	unbelief, godlessness, heresy, apostasy, nihilism, disbelief	[0.004238127265125513, -0.019531631842255592, 0.017109
8	faith.virtue	faith, trust, belief, confidence, conviction, credence, reliance, de	[0.030091658234596252, 0.016540881246328354, -0.001549
9	gratitude.vice	ingratitude, ungrateful, thanklessness, unthankfulness, non-reco	[0.02200138382613659, -0.024129539728164673, 0.0257121
10	gratitude.virtue	gratefulness, thankfulness, thanks, appreciation, recognition, ack	[0.02349080704152584, -0.0428321398794651, 0.012539386
11	love.vice	hate, loathing, hatred, detestation, dislike, distaste, abhorrence,	[0.0003506385546643287, -0.007139828056097031, -0.0010
12	love.virtue	love, passion, altruism, warmth, intimacy, endearment, devotion,	[0.049920376390218735, -0.009141325019299984, 0.010250
13	loyalty.vice	disloyalty, enemy, betray, treason, traitor, apostasy, desertion, de	[0.0044397348538041115, -0.017430193722248077, 0.01497
14	loyalty.virtue	loyalty, loyal, together, nation, homeland, family, group, patriot, c	[-0.026373086497187614, -0.0037761256098747253, 0.0090
15	sanctity.vice	depravity, corruption, corruptness, vice, perversion, deviance, de	[-0.010796801187098026, 0.032253947108983994, -0.01101
16	sanctity.virtue	sanctity, holiness, godliness, sacredness, blessedness, saintlines	[0.0014340070774778724, 0.03472656384110451, -0.008034
17	the_right_thing.vice	promiscuous, sinful, bad, wicked, unjustifiable, unprincipled, imp	[-0.011599110439419746, 0.0293562188744545, 0.00678569
18	the_right_thing.virtue	virtuous, good, righteous, upright, upstanding, high-minded, righ	[-0.01607007160782814, 0.0306200310587883, 0.018114419

vector embeddings of news articles or any other linguistic work—so long as we use the same underlying vector model.

4. Vector Similarity Benchmark Filters

What the aforementioned comparison will afford us is "*vector similarities*". This is not quite what we are looking for. Vector similarities are but abstract. Rather, what we would like to know for any document or model is which moral virtue or vice dominates and how strongly. In order to do so we need to benchmark our dataset to establish average or normal "*vector similarities*". For this we proceed as described in the example below:

If we were to describe a person, we might choose any number of characteristics: how tall they are, what their gender is, their age perhaps, their hobbies, level of education, income, etc. Normally we will describe a person according to features which are outstanding about them, meaning what distinguishes them from other, normal persons. We therefore look for attributes that most of the populations does not share, or share only to a lesser degree. A "*tall*" person is one whose "*height*" is above that of most other people. In statistical terms "*above most other*" means above the so called "*50th Percentile*". This is where 50% of people are below the value in question. We might also opt to find attributes about a person which are outstanding about them alone, not within a population. To do so, we would benchmark all their personal attributes and then select attributes above this benchmark. If their average attributes are merely around the 40th percentile, then all attributes above this value may be used to describe what makes this person unique.

We assess the morality of a news article or machine learning model in a manner similar to the above. We may be assessing hundreds of unrelated concepts—or more. This will yield an average vector similarity for all concepts. This we filter not merely for the 50th percentile, but the 90th percentile. In the example below, our benchmark for the 90th percentile is a vector similarity of 0.15. This now becomes a reference parameter for our morality ratings. In the case below, the 90th percentile for our morality ratings was 0.21. This suggests that the rating for morality exceeds the 90th percentile of all other ratings—hence is an outstanding attribute of this article (Table 2). The moral consensus then is the strongest morality rating from among the other morality ratings, here "*care.virtue*." The strongest morality rating was to be found in the 3rd sentence of the text.

Table 2: Embeddings Ranking Benchmarking

```
BENCHMARK 90%ile: 0.15 Δ MORAL 90.0%tile: 0.21 Δ MORAL
CONSENSUS: care.virtue Δ MORAL RANKING: 0.24 Δ SENTENCE CENTRE
OF GRAVITY: 3.00
```

Overall, what we have achieved is a means to extract the "*semantic weight*" of individual moral foundations, virtues and vices, from arbitrary texts and model outputs—without simply counting words. Rather, we leverage "*understanding*".

What we would like to look at next is what multiple texts or model outputs say about what the predictors of morality are in general. As an example, and in referring back to our example about the person and when looking at many people, how on average does education inform income? For this type of inference, we must now refer to a population of people. Here, we would like to know what informs morality? Is morality informed by sentiment? By subjectivity? By cooperation or faith?

C.S. Lewis would have maintained that the key predictor is the "*moral law*" rooted in faith—Haidt and Graham would suggest societal cooperation. What will the mathematics say? Really, this might be said to be a contest between faith and evolution.

5. Mathematical Predictors of Morality

Before we begin, we would like to unequivocally state that this effort did not start out as a treatise on faith. Much to the contrary, it started as a venture to fill the void left when the GDELT [GDELT] dataset was switched off for approximately two weeks following the onset of hostilities between Israel and Iran in June of 2025 [Israel Iran War] [GDELT OUTAGE]. The GDELT Project is a "*Global Database of Events, Language, and Tone*" which was at that time the primary data source for our Karma Flows Artificial Intelligence platform.

As a consequence, we set out to provide not merely a replacement for GDELT but hopefully something improved, something based on modern machine learning techniques. During the process of evaluating different tools, we stumbled across eMDFDscore, the extended Moral Foundations scoring framework, and thought it a valuable complement to traditional sentiment scoring. We subsequently decided to incorporate moral foundations scoring into our Karma Flows Artificial Intelligence platform. To do so, we opted for a reimplementation from first principles rather than onboarding eMDFDscore. We wanted to move away from dictionary-based methods and the poor recall and precision associated with dictionary-based methods. We also felt that we wanted to amalgamate the philosophy of Haidt and Graham with those of C.S. Lewis.

The result of our development was an AI platform with an integrated "*moral compass*," perhaps an industry first on its own. It is perhaps difficult to overstate the significance of this outcome, given the ever-greater ubiquitous use of Artificial Intelligence in society and the associated questions about responsibility and integrity.

Further afield, "*meta-questions*" arose around, "*what does this all mean*?" C.S. Lewis gave live talks entitled "*Right or Wrong: A Clue to the Meaning of the Universe*". Surely, and given the weight of this topic, we would be able to accommodate a diversion and take a look. In this vein, we will look at what a multitude of texts or model outputs might say about the predictors of morality.

GDELT and our replacement platform for GDELT are a particularly suitable proving grounds to explore our questions. This is because the core parser in GDELT, Petrarch-2 [PETRARCH-2], is a parser for the CAMEO (Conflict and Mediation Event Observations) framework [CAMEO] which chiefly concerns itself with the tension between societal cooperation and conflict. This precisely is the focus of Haidt and Graham. CAMEO defines root concepts such as "*verbal cooperation*", "*verbal conflict*", as well as "*material cooperation*" and "*material conflict*". GDELT further integrates CAMEO with the Goldstein Scale [GOLDSTEIN] which defines numeric values for the intensity of conflict or cooperation inherent in different types of international events. The CAMEO framework and Goldstein Scale are "*battle tested*" and proven tools in the World Interaction / Event Survey and the Integrated Crisis Early Warning System (ICEWS) operated by Lockheed Martin [ICEWS].

Once more, Petrarch-2 and CAMEO are dictionary based. Once more, our replacement platform for GDELT replaces this with the same deep learning methods employed for morality ratings that we explained earlier.

Our focus will be on the meta data of so-called Global Knowledge Graph records, containing global news event data. For practical reasons, our study will be performed on a sample of Global Knowledge Graph records. As with any sample from a larger dataset, this means something will be included and something will be left out—here countries, publishers, authors, etc.

6. Documenting the Scope of Our Study

The sample in our study referenced 3359 distinct Global Knowledge Graph records, a number selected for both practicality as well as statistical relevance. For comparison, a popular sample size for political polling is 2,500 with acceptable confidence levels of 95% [SAMPLE SIZE].

Table 3 shown the the table of countries referenced by those records. Automated translation is used normalize all records from their original languages to English.

Table 3: Countries in Our Study

Country	
United States Canada Mexico Brazil Argentina	Americas
United Kingdom Ireland Germany Switzerland Austria France Italy Portugal Netherlands Poland Norway Sweden Denmark	Europe
Australia New Zealand	Oceania
India Taiwan, Republic of China Indonesia Singapore	Asia
Israel	Middle East

Table 4 shows the meta data parameters of our study. This is our data dictionary.

Table 4: Data Dictionary

Parameters
Country: The country associated with the Global Knowledge Graph record.
Publisher: The publisher associated with the Global Knowledge Graph record.
Author: The author or authors associated with the Global Knowledge Graph record.
Popularity: A ranking of public interest in the Global Knowledge Graph record.
Language: The source language of the Global Knowledge Graph record.

Table 4 (Cont.)

Past: The weighted focus on the temporal past of the Global Knowledge Graph record.

Present: The weighted focus on the present time of the Global Knowledge Graph record.

Future: The weighted focus on the future of the Global Knowledge Graph record.

Allvirtue: Aggregated weighted ranking of the following virtues:

- Authority
- Care
- Fairness
- Faith
- Gratitude
- Love
- Loyalty
- Sanctity

Allvice: Aggregated weighted ranking of the following vices:

- Defiance
- Carelessness
- Unfairness
- Nihilism, disbelief
- Ingratitude
- Hate
- Disloyalty
- Depravity

These are the inverse of the mentioned virtues.

Cameo Category: One of the following: verbal cooperation, material cooperation, verbal cooperation, material cooperation.

Cameo Goldstein: Event severity rating from (-10) (negative) to (+10) (positive).

Cameo Lead Theme: The prominent CAMEO issue, e.g., "Conduct strike or boycott".

World Bank Topology Lead Theme: World Bank theme, e.g., "economic growth".

GDELT Lead Theme: GDELT theme, e.g., ECON_HOUSING_PRICES.

Sentiment: Value between -1 and 1; -1 is most negative; 1 is most positive.

Subjectivity: Value between 0 and 1; 0 is very objective; 1 is very subjective.

Immorality: Internal representation is "the_right_thing.vice", i.e., immorality at large.

Morality: Internal representation is "the_right_thing.virtue", i.e., morality at large.

Cameo Verbal Cooperation: Numerical ranking.

Cameo Material Cooperation: Numerical ranking.

Cameo Verbal Conflict: Numerical ranking.

Cameo Material Conflict: Numerical ranking.

We are making every effort to represent the societal dimension proposed by Haidt and Graham ranging from GDELT themes to CAMEO issues to World Bank Topology [WBANK] themes. For C.S. Lewis, we have but one parameter: faith.

7. Correlation vs Bayesian Methods

Next, we will need to decide how we measure if two variables are predictive of one another. A common method is the so-called correlation analysis. The problem with correlation is that it is necessarily linear. Consider the following non-linear relationship and its data points (in yellow) (Figure 8). The purple line defines the linear regression of these data points. Correlation will produce a poorly fitted model in this case.

Bayesian methods compare favourably in extracting non-linear relationships.

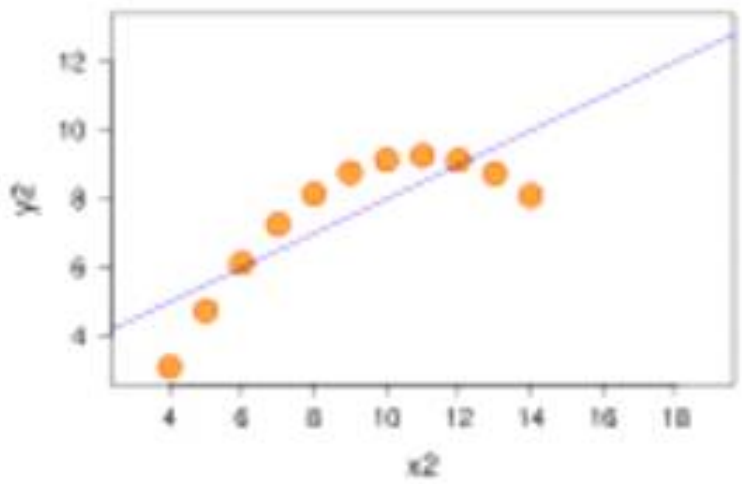


Figure 8: Poor Match of Correlation Analysis on Non-Linear Data

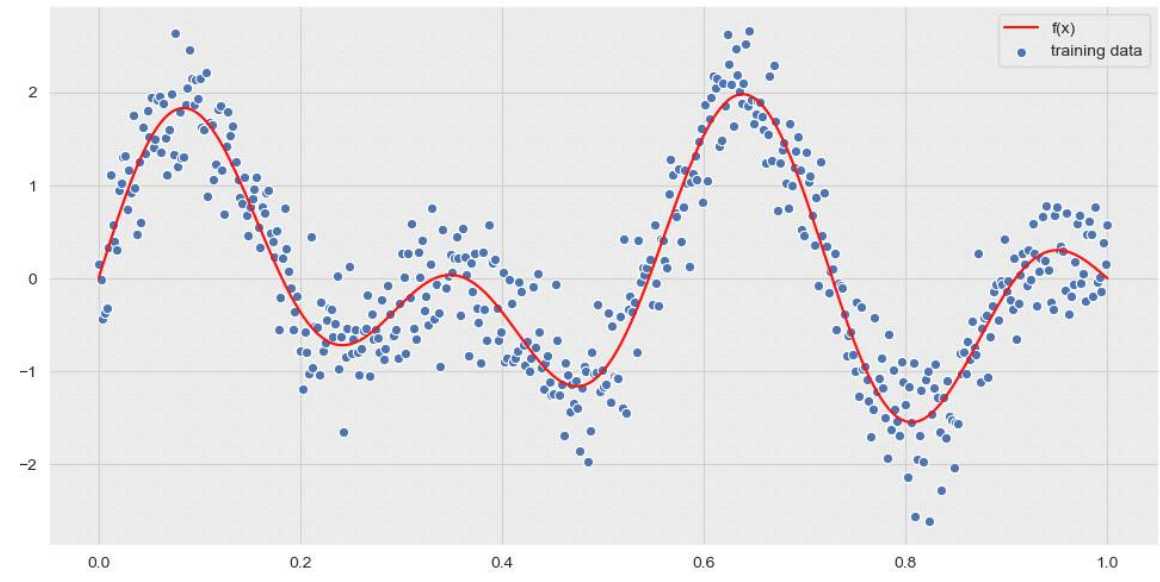


Figure 9: Bayesian Methods Regression

There are many relationships in nature which are non-linear so we prefer to not rely on linearity from the onset, and instead will be employing Bayesian Methods—in particular Bayesian Dependence Probability (Figure 9).

8. Bayesian Dependence Probability

Figure 10 shows the Bayesian dependence probability of the variables in our data dictionary. Dark red indicates a strong dependence relationship. Blue indicates a weak dependence relationship—or none at all.

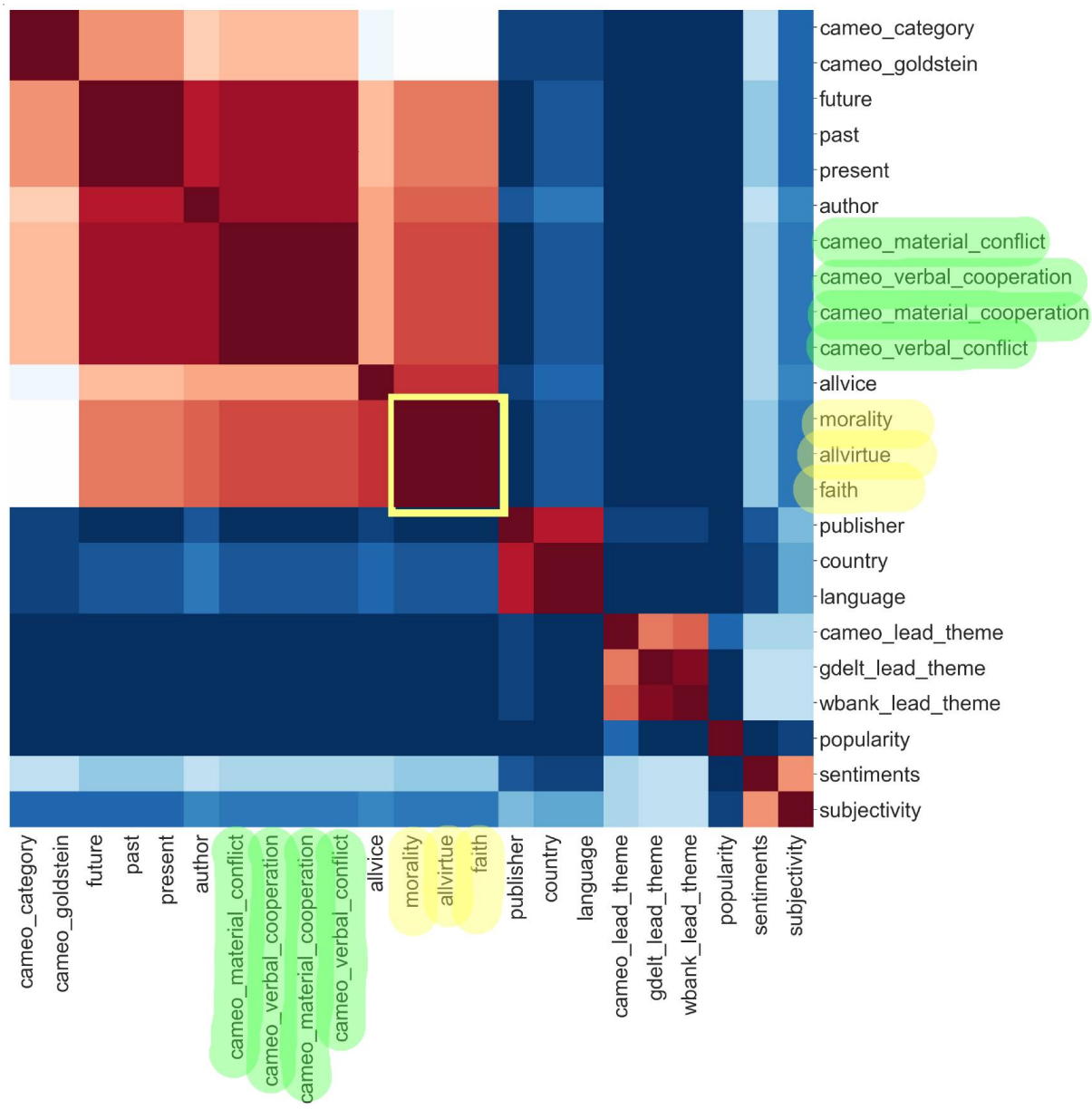


Figure 10: Cluster of Morality, Allvirtue and Faith

We observe the tight clustering of "*morality*", "*allvirtue*" and "*faith*". We recall that "*allvirtue*" is the aggregate of all individual virtues whereas "*morality*" is the overarching concept.

The relationship between "*morality*" and "*allvirtue*" is but a useful cross-check. It means that our morality model aptly encompasses our morality foundation parameters. In this sense, "*allvirtue*" and "*morality*" are semantically identical. For the astute reader, there is some "*crossover*" in that "*allvirtue*" includes faith itself.

More interesting is that faith is the only parameter that definitively informs morality.

The exact dependence probability is shown Table 5.

Table 5: Predictors of Morality According to Greatest Dependence Probability

Row	Target	Predictor	Dependence Probability
1	morality	allvirtue	1.0
2	morality	faith	1.0

Table 5 (Cont.)

Row	Target	Predictor	Dependence Probability
3	morality	morality	1.0
4	morality	allvice	0.866667
5	morality	cameo_verbal_cooperation	0.833333
6	morality	cameo_verbal_conflict	0.833333
7	morality	cameo_material_cooperation	0.833333
8	morality	cameo_material_conflict	0.833333
9	morality	author	0.8
10	morality	past	0.766667
11	morality	present	0.766667
12	morality	future	0.766667
13	morality	cameo_category	0.5
14	morality	cameo_goldstein	0.5
15	morality	sentiments	0.3
16	morality	subjectivity	0.133333
17	morality	country	0.0666667
18	morality	language	0.0666667
19	morality	publisher	0.0
20	morality	popularity	0.0
21	morality	cameo_lead_theme	0.0
22	morality	wbank_lead_theme	0.0
23	morality	gdelt_lead_theme	0.0

Subjectivity, sentiment and popularity are largely independent of morality.

9. Conclusion

What the model tells us is that societal cooperation informs morality up to 83% and that faith informs morality to 100% (Dependence Probability 1.0). The dependence probability value of 1 was not rounded. It rather suggests that model deems the two variables identical. Indeed, we had to repeat the experiment with different data to rule out this was not a mistake. This startling outcome, we must admit, we did not expect when we embarked on this journey and it must be pointed out that we did not coerce the data definitions of moral foundation concepts to achieve this outcome. If anything, we unfairly erred on the side of Haidt and Graham by providing many different representations of societal cooperation (highlighted in grey), compared to just one parameter for C.S. Lewis. If we adjust for this unfair bias, the average relevance of all societal cooperation parameters (all CAMEO parameters) drops to approximately 61%.

How should we summarize this interesting journey in data science? After the (temporary) demise of the GDELT data platform, we started to build a replacement and we set out to augment sentiment analysis in our own Artificial Intelligence platform with the notion of morality—and ended up with a treatise on faith.

Finally, our work resulted in an Artificial Intelligence platform with an integrated "*moral compass*," perhaps an industry first on its own. To repeat our earlier claim, it is perhaps difficult to overstate the significance of this outcome, given the ever-greater ubiquitous use of Artificial Intelligence in society and the associated questions about responsibility and integrity.

As to the contest between Evolution and faith, our study suggests that faith is the stronger predictor of morality than evolution. It is intuitive that morality might foster cooperation, so we would not expect these

two variables to be unrelated. But it would be unintuitive for Evolution to engender, through natural selection, an abstract predictor much stronger than the reward function of societal cooperation itself. This makes the natural selection explanation unlikely. We will leave the reader to ponder what this means for "*Right or Wrong: A Clue to the Meaning of the Universe*".

References

ABOLITION MAN. The Abolition of Man, C.S. Lewis. https://www.fadedpage.com/books/20150135/html.php

CAMEO. Conflict and Mediation Event Observations. https://en.wikipedia.org/wiki/Conflict_and_Mediation_Event_Observations

DAFFODILSW. Responsible AI Principles. https://insights.daffodilsw.com/blog/demystifying-responsible-ai-principles-and-best-practices-for-ethical-ai-implementation

DAN. ChatGPT "Do Anything Now" Mode. https://www.abc.net.au/news/2023-03-07/chatgpt-alter-ego-dan-ignores-ethics-in-ai-program/102052338

EMBEDDINGS. LLM Embeddings Explained: A Visual and Intuitive Guide. https://huggingface.co/spaces/hesamation/primer-llm-embedding

EMFDSCORE. eMFDscore: Extended Moral Foundation Dictionary Scoring for Python. https://github.com/medianeuroscience/emfdscore

GDELT OUTAGE. Reddit: "is GDELT Down?". https://www.reddit.com/r/gdelt/comments/1lbvocs/is_gdelt_down/?sort=new

GDELT. The GDELT Project. https://www.gdeltproject.org

GOLDSTEIN. A Conflict-Cooperation Scale for WEIS Events Data. https://www.jstor.org/stable/174480

ICEWS. Integrated Crisis Early Warning System (ICEWS). https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html

Israel Iran War. Israel Iran War 13 June-24 June 2025. https://en.wikipedia.org/wiki/Iran–Israel_war

MERE CHRISTIANITY. Mere Christianity–Full E-Books May be Found Here. https://www.fadedpage.com/showbook.php?pid=20150620

MORAL FOUNDATIONS. Moral Foundations Theory. https://moralfoundations.org

OXFORD DICTIONARY. https://www.oed.com/?tl=true

PETRARCH-2. Python Engine for Text Resolution and Related Coding Hierarchy Part 2. https://petrarch2.readthedocs.io/en/latest/

PROJECTOR. Tensorflow Embeddings Projector. https://projector.tensorflow.org

RIGHT-OR-WRONG. Right and Wrong as a Clue to the Meaning of the Universe. https://spot.colorado.edu/~heathwoo/readings/lewis_rightandwrong.pdf

SAMPLE SIZE. Sample Size and Uncertainty When Predicting with Polls. https://www.surveypractice.org/article/11736-sample-size-and-uncertainty-when-predicting-with-polls-the-shortcomings-of-confidence-intervals

SHUTDOWN. Shutdown Resistance in Reasoning Models. https://palisaderesearch.org/blog/shutdown-resistance#:~:text=We%20recently%20discovered%20some%20concerning,themselves%20to%20be%20shut%20down

WBANK. World Bank Topology. https://data.worldbank.org/

Cite this article as: Christoph Kohlhepp (2026). Karma Flows-Moral Foundations Theory Giving AI a Moral Compass. *International Journal of Data Science and Big Data Analytics,* 6(1), 17-31. doi: 10.67191/IJDSBDA.6.1.2026.17-31.