Rodrigue Castro Gbedomon
Chemin de Castelver 4, 1255 Veyrier Suisse
Rodrigue.gbedomon@unige.ch

Data Science Project

Are public, policy, and science aligned on human–nature relationships?

Conceptual Design Report

04 October 2025

Abstract

This project investigates the normative positions on human-nature relationships, their prevalence and share across societal areas (public opinion, policy agenda and science debate) and cultural contexts. Building on an existing fuzzy matrix of 71 normative positions and 25 worldviews, the project aims to propose an typology of normative positions using unsupervised machine learning and to explore how this typology is reflected across societal areas and cultural contexts. In the first phase, we will use the clustering techniques to identify coherent groups of normative positions and extract their key defining features. In the second phase, a culturomics approach will be applied to digital cultural traces (including Google Trends, Wikipedia, GDELT, policy records, and scientific corpus) to analyze how these normative positions circulate across societal areas, cultural contexts and over time.

Table of Contents

1 Project Objectives

Humans perceive and interact with nature in diverse, sometimes conflicting ways, reflecting multiple conceptions of the human–nature relationship [1-3]. These interactions and perceptions can vary over time, across cultures and under different economic conditions, underpinning plural normative positions in nature conservation. Therefore, the debate on nature conservation is value driven and highly polarized, challenging inclusive actions towards the future we want. In a recent publication, Gbedomon et co-authors [4] identified five normative positions in conservation science (Human dominion, one with nature, just another species, nature protection, green economy), each, each reflecting distinct values and worldviews about the human–nature relationship. The typology is largely built on expert judgment, making it susceptible to bias. The underlying matrix is over-specified, and several dimensions lack clear relevance or construct validity. In addition, it remains unclear whether, and to what extent, this typology is reflected in public opinion, policy agendas, and scientific debates, as well as how it captures differences across these societal areas and cultural contexts. These issues compromise transparency and render the work difficult to replicate.

The aim of this project is to refine and improve the existing typology of normative positions on human–nature relationships using unsupervised matching learning and culturomics approaches. In a first step, we will use matching learning to evidence similar group and extract the key features of each group of normative positions. Second, we will examine how these positions are expressed across public attention, policy agendas, and scientific debates, and how they vary across cultural contexts and over time. Finally, the project seeks to determine the extent of alignment or divergence between societal arenas, providing a more inclusive and globally representative understanding of the value systems underpinning conservation discourses.

2 Methods

Infrastructure and tools

For this project, we will combine local resources with cloud-based services for data storage and processing. The initial and processed datasets will be stored on Google Drive but will also be available on Github for versioning and collaboration. For data processing, we will use Google Colab to run python.

Software Libraries

All analyses will be conducted primarily in Python 3.11. Core libraries identified so far as useful in this study are included in table 1.

Table 1. Libraries to be used in the study

Group	Libraries	Functions
data Handling	pandas	data manipulation, matrix handling, and preprocessing.
	numpy	efficient numerical operations and linear algebra.
	openpyxl	reading the Excel matrix and managing structured inputs.
	scipy	distance computations, clustering metrics, and statistical utilities.
Machine Learning & Statistical Analysis	scikit-learn	Preprocessing (standardization, scaling, encoding), Dimensionality reduction (PCA, t-SNE), Clustering (K-Means, Agglomerative, Spectral, Gaussian Mixtures), Model evaluation (silhouette score, Davies–Bouldin, etc.)
	umap-learn	for advanced non-linear dimensionality reduction (UMAP) to reveal structure in fuzzy worldview data
	hdbscan	for density-based clustering that can handle variable cluster shapes and noise (important for fuzzy categorical structures)
	Yellowbrick	visual diagnostics for clustering (elbow plots, silhouette visualization)
Feature Importance & Interpretation	scikit-learn linear models	predict cluster membership and extract key features per group.
Visualization & Exploration	matplotlib	base plotting for embeddings, dendrograms, and cluster signatures
	seaborn	for statistical heatmaps, pairplots, and stylistic enhancements

Statistical Methods

Phase 1. Reconstruction of the typology of normative positions

Descriptive statistics and exploratory analyses (frequency distributions, correlation matrices, and hierarchical heatmaps) will be performed to examine the overall structure of the data and assess variability across values. To reduce dimensionality and uncover latent structures, Principal Component Analysis (PCA) will be applied, with the number of retained components determined using eigenvalue criteria and scree plots. In addition, non-linear dimensionality reduction methods such as Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) will be employed to visualize the structure of the data in lower dimensions while preserving neighborhood relationships.

Unsupervised clustering techniques will then be applied to identify groups of normative positions with similar values. K-means clustering will be implemented on PCA-transformed data, with the optimal number of clusters determined through elbow plots, silhouette coefficients. Agglomerative hierarchical clustering using Ward's linkage will provide complementary insights into hierarchical relationships between worldviews, while HDBSCAN will be used to detect clusters of varying shape and density and to identify potential outliers. Once clusters are established, each group will be characterized by computing cluster centroids and standardized mean scores for each worldview ou value. Group differences across clusters will be assessed (using one-way ANOVA or Kruskal–Wallis tests) depending on distributional assumptions. In parallel, multinomial logistic regression will be trained to predict cluster membership based on worldview scores. Model coefficients values will be examined to identify key discriminant features defining each normative group.

Phase 2. Assessing the prevalence and alignment of the normative positions over time and across societal areas and cultural contexts

Our primary task is normative positions inference and alignment measurement. We curate seed lexicons and exemplar snippets for the typology of normative positions (from the STEP1), generate silver labels via weak supervision, and fine-tune multilingual transformers to obtain calibrated probabilities. Performance is reported per language and societal areas, with error analysis uses confusion matrices and token-level SHAP. We aggregate positions distributions monthly by language and country, quantify polarization with Shannon entropy and Gini–Simpson, and assess alignment across societal areas using Jensen–Shannon divergence.

Software Ecosystem and Reproducibility

All analyses will be conducted primarily in Python. All the project will be maintained in GitHub, with the relevant documentation including notebooks, README, a data dictionary, labeling guide, etc. The other resources (code, cleaned data, metadata) will be made openly available but with prior requests in line with FAIR principles.

3 Data

Existing data and data collection

For this study we have already preliminar dataset on normative positions on human-nature relationships. This dataset was generated through peer-reviewed literature, grey literature, and expert opinions. We have conducted a systematic search in Web of Science Core Collection (French/English) using keywords on nature (e.g., biodiversity, ecosystem, conservation) and on normative orientations (e.g., worldviews, paradigms, narratives, discourses, values). From 6,369 initial records, 5,754 were excluded by title/abstract screening. The remaining 615 were complemented with 35 items (including grey literature) from authors' bibliographies, yielding a final corpus of 650 documents describing how human–nature relations are conceptualized, measured, or valued. From the 650-paper universe, we identified 71 normative positions.

In addition to this existing dataset, we will assemble a multilingual corpus of digital cultural traces across the three social areas : Public, policy and science

Public

- Google Trends (web search interest). We will pull keyword series per language–country to capture *salience* and *event-driven spikes* in public attention to human–nature concepts and positions proxies. Where appropriate, we will supplement normalized indices with available absolute-volume estimates (documented sensitivity analysis).
- Wikipedia Pageviews (Wikimedia). Pageview time series for concept pages provide a second, behaviorally distinct proxy of *sustained curiosity/learning*, useful for triangulating Google search patterns and estimating attention.

Policy agenda

- GDELT v2 (Global Database of Events, Language, and Tone). We will query multilingual web/broadcast news to obtain *coverage volume*, *geolocation*, and *tone* around our concept and stance lexicons, aggregated to country–month. This captures the media agenda that shapes and reflects policy discourse.
- Parliamentary records & international policy briefs (e.g., UN Official Document System; IPBES publications library). Where accessible, we will scrape/download debates, bills, and briefs to anchor media signals in official policy framing and vocabulary.

Scientific knowledge

Crossref REST API and Scopus metadata. Titles/abstracts of peer-reviewed articles and proceedings indexed under conservation-relevant subject areas provide the research supply lens.

Data cleaning and processing

The preliminarnar dataset was cleaned and exist as a fuzzy matrix. A fuzzy variable (normative position) is scored as a=(x1, x2, x3) for each value-state. For instance, codes (1, 0, 0) (0, 1, 0) and (0, 0, 1) indicate that the normative positions held a "Pros", "Cons" and "Neutral" position, respectively, for the targeted value-state. The absence of information data is thus denoted (0, 0, 0). As such, the fuzzy coding approach can accommodate for uncertainty and conflicting sources of information within the ordination analysis [5, 6].

Before proceed the analyses, the fuzzy worldview matrix will be transformed into a numerical format by assigning +1 to “Pros,” −1 to “Cons,” and 0 to “Neutral,” followed by standardization to zero mean and unit variance to ensure comparability across variables.
As for the data to be collected, we will check for missing data using simple summaries and charts.

Overview of the existing data

Here is an overview of the preliminar dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72 entries, 0 to 71
Data columns (total 26 columns):
 #   Column                           Non-Null Count  Dtype 
---  ------                           --------------  ----- 
 0   Normative positions              72 non-null     object
 1   Human exceptionalism             72 non-null     int64 
 2   Naturalized human distinctiveness 72 non-null     int64 
 3   Total naturalism                 72 non-null     int64 
 4   Dominion                         72 non-null     int64 
 5   Stewarship/Responsability        72 non-null     int64 
 6   equality                         72 non-null     int64 
 7   Reverence for nature             72 non-null     int64 
 8   Pristine nature                  72 non-null     int64 
 9   Nature with some alteration      72 non-null     int64 
 10  Domestic nature                  72 non-null     int64 
 11  Biophobia                        72 non-null     int64 
 12  Biophilia                        72 non-null     int64 
 13  Unconcerned                      72 non-null     int64 
 14  Intrinsic                        72 non-null     int64 
 15  Instrumental                     72 non-null     int64 
 16  Relational                       72 non-null     int64 
 17  No intervention                  72 non-null     int64 
 18  Light intervention               72 non-null     int64 
 19  Directed intervention            72 non-null     int64 
 20  Biotechnology                    72 non-null     int64 
 21  Technical                        72 non-null     int64 
 22  Scientific (Academic)            72 non-null     int64 
 23  Indigenous                       72 non-null     int64 
 24  democratic (local)               72 non-null     int64 
 25  Economic (market)                72 non-null     int64 
dtypes: int64(25), object(1)
memory usage: 14.8+ KB
None
```

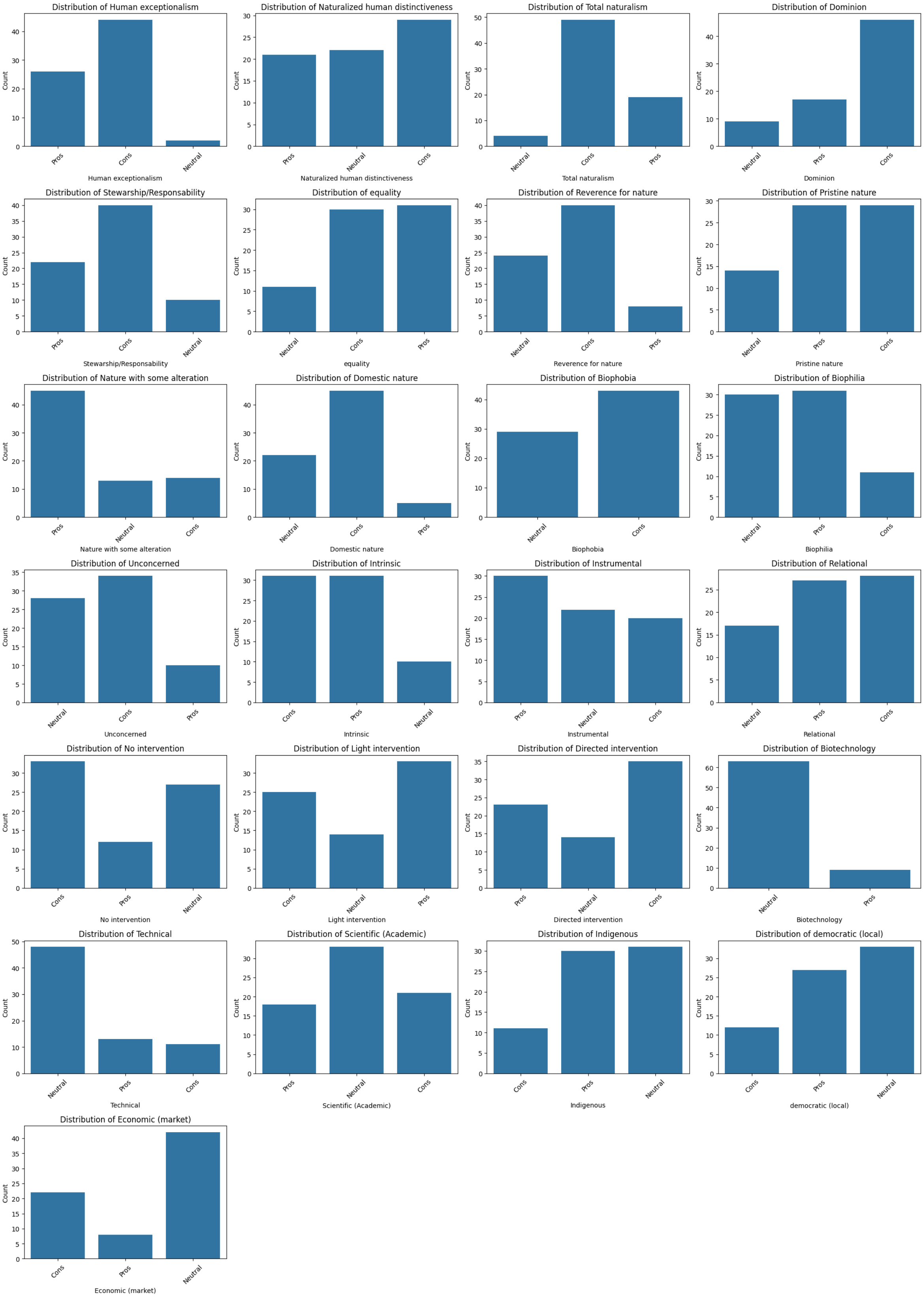

Figure 1. Distribution of the worldviews

4 Metadata

We will be using different datasets as described in the previous sections. For these datasets, the follwing metadata will be procuded and made available.

- First, data provenance metadata including the original sources, date of extraction, language, and geographic scope.
- Second, processing metadata will describe all pre-processing steps, such as text cleaning, transformation, tokenization, embedding methods, and parameter values used in normative positions classification or clustering.
- Third, analytical metadata will include the versions of software, libraries, and models employed, as well as details of statistical tests and model configurations.

All metadata will be stored in structured formats alongside the datasets and analysis pipelines and maintained within the project's version-controlled repository on GitHub.

5 Data Quality

We have already performed quality checks on the existing dataset. For the remaining datasets we are at the initial stage of data collection, focusing on exploring databases and variables. The literature [7] highlights three main types of data-related risks : (1) data volume, (2) data noise or lack of precision, and (3) missing or incomplete data. To address these challenges and ensure high data quality, we have established the following guiding principles:

- Data volume: For time series, we will collect continuous records spanning 25 years for Google Trends and 9 years for GDELT. For text corpora (Scopus abstracts, policy briefs), we aim for at least several thousand documents per language, which is necessary to enable meaningful clustering and robust analyses.
- Data precision and resolution: Queries and keywords must be identifiable at a fine-grained level (by language, region, and time), and sentiment/tone measures should yield consistent and comparable scores across datasets.
- Data completeness and consistency: Excessive gaps in search volumes or news coverage would undermine the tracking of long-term trends. Therefore, minimizing missing values is essential. Furthermore, multilingual consistency is critical, as cross-linguistic comparability underpins the cross-cultural validation of normative positions.

Third, issues of completeness and missing values are critical: excessive gaps in search volumes or news coverage would limit the ability to track long-term trends. Finally, consistency across languages is required, since multilingual comparability underpins the cross-cultural validation of normative positions.

6 Data Flow

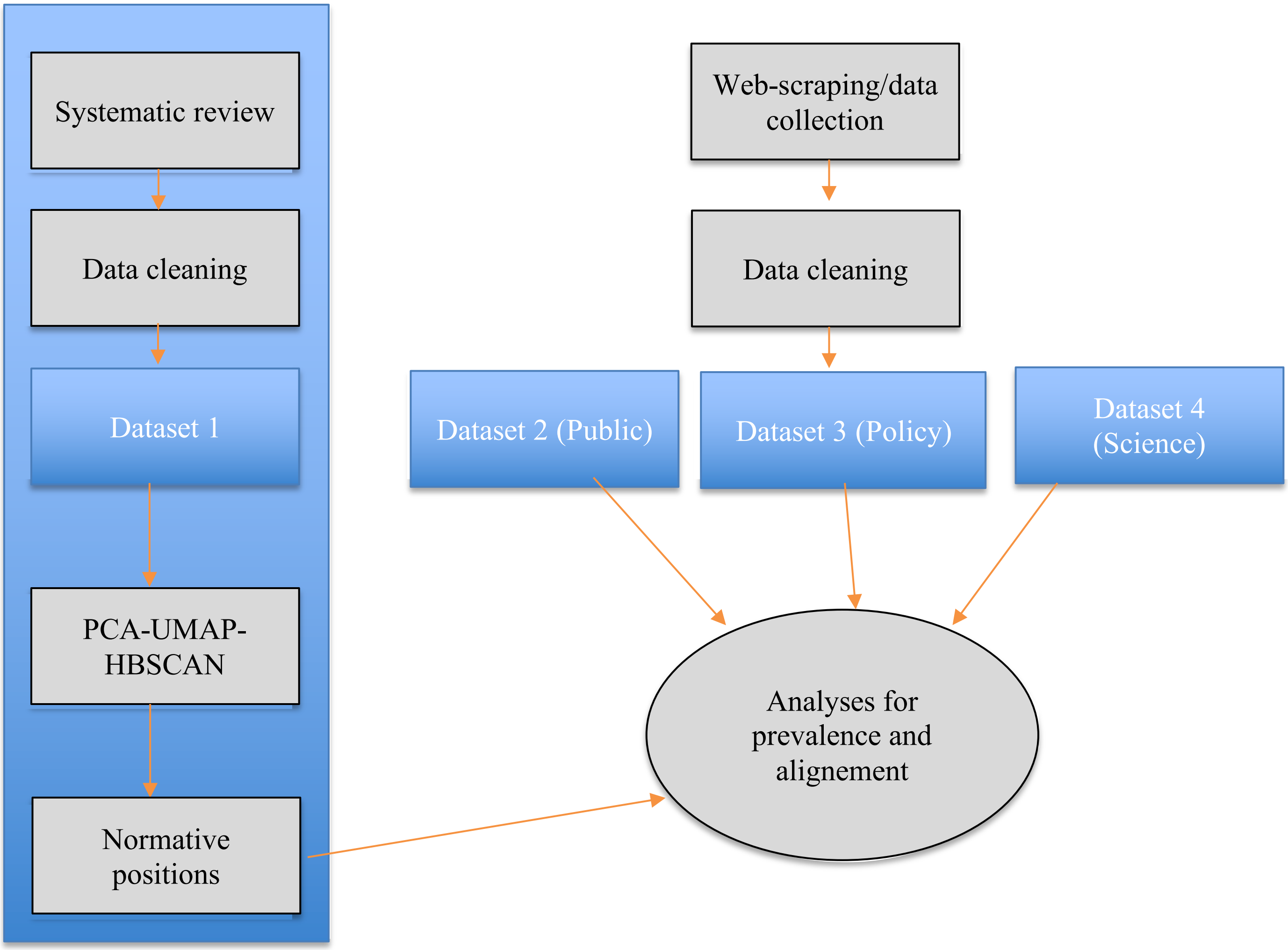

Figure 2: the workflow of the project.

7 Data Model

Conceptual level

At the conceptual level, the data model is grounded in the need to provide stakeholders in nature conservation with an integrated and inclusive taxonomy of normative positions, as well as a clear understanding of how these positions evolve over time, across societal domains, and within different cultural contexts. Developing such a taxonomy is crucial for elucidating differences in perceptions of human–nature relationships and for fostering more inclusive and context-sensitive conservation strategies. By making these normative stances explicit, the study aims to help conservation actors become more aware of their own value systems and better equipped to provide guidance that reflects and accommodates a diversity of perspectives.

Logical level

At the logical level, the data model focuses on the following entities :

- Normative positions. Each position represents a distinct posture or conceptualization of human-nature relationships.
- Worldviews or values. Analytical dimensions of normative positions. Worlviews refers to a collection of principles (for example, based on ethical frameworks), preferences (such as religious beliefs), and choices regarding how the principles are applied.

In this project we try to link the two abovementioned entities and create clusters of normative positions on human-nature relationships. In addition we try to see whether and to what extent these clusters resonate with societal areas. This model is important to evidence the plurality of normative positions on human-nature relationships, and help people be aware of their positions.

Physical level

Given the scope of the work, we do not anticipate heavy computational demands. Google Drive will be sufficient for storing raw and processed data, and Google Colab will suffice for data cleaning, exploration, and basic analyses. We therefore do not expect to require high-performance computing, dedicated servers, or additional tools beyond the standard Python stack available in Colab. If data volume grows substantially (especially during the additional data collection phase), we will reassess resource needs.

8 Documentation

The project will be documented to ensure transparency, reproducibility, and accessibility. First the datasets used will be accompanied by detailed metadata and data dictionaries describing sources, structure, processing steps, and versioning. The analytical workflows will be documented version-controlled repersitory on Github. The code will also be maintained in a Github.

9 Risks

The main risk for this project is associated with the culturomics approach planned for the second phase. Accessing digital cultural data may prove challenging due to platform policies, paywalls, and API restrictions. For example, the free version of Google Trends limits users to a very small number of monthly requests, which may constrain the scope of data collection. In addition, historical data are often incomplete, geographically biased, or unevenly distributed across languages, which can create significant gaps and limit the validity of cross-cultural comparisons. To mitigate these risks, the project will diversify data sources, archive data early and systematically, and remain flexible in adapting methods or concepts if certain sources become inaccessible.

10 Preliminary Studies

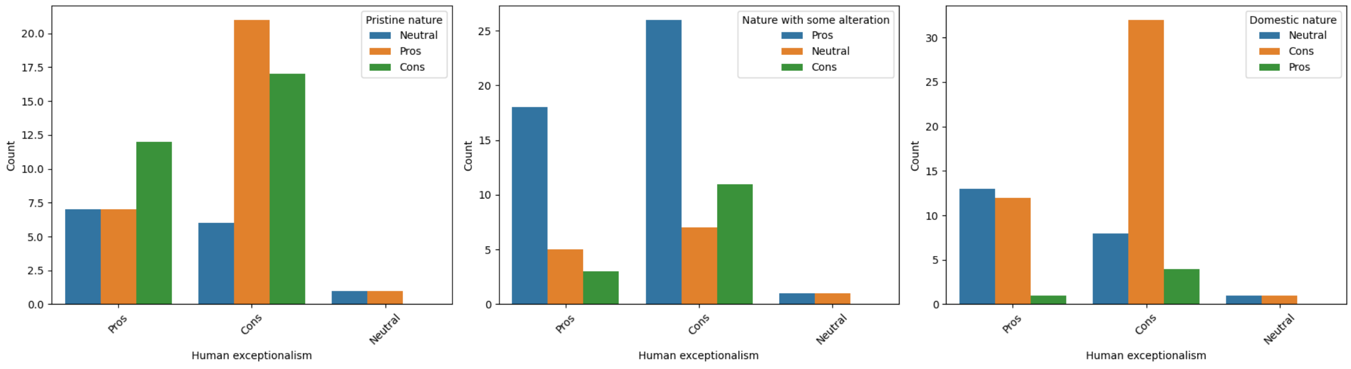

Figure 3. Relationships between Human exceptionalism and Pristine nature, betweeb Human exceptionalism and Nature with some alteration, and between Human exceptionalism and Domestic nature

11 Conclusions

This project represents an innovative contribution to conservation science by combining unsupervised machine learning with culturomic analyses to map how diverse value systems shape human–nature relationships. The methodological framework will generate a more nuanced, globally informed understanding of normative positions across public, policy, and scientific domains. The resulting taxonomy can help conservation actors become more aware of their own value positions and design strategies that are culturally inclusive and responsive to societal diversity.

Acknowledgements

We would like to thank the co-authors of the previous study for approving the use of the initial dataset on normative positions and worldviews.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: October 04th 2025 Signature(s):

References and Bibliography

[1]. Chapman, M., T. Satterfield, and K.M.A. Chan, How value conflicts infected the science of riparian restoration for endangered salmon habitat in America's Pacific Northwest: Lessons for the application of conservation science to policy. Biological Conservation, 2020. 244.

[2]. Pascual, U., et al., Valuing nature's contributions to people: the IPBES approach. Current Opinion in Environmental Sustainability, 2017. 26-27: p. 7-16.

[3]. Wienhues, A., L. Luuppala, and A. Deplazes-Zemp, The moral landscape of biological conservation: Understanding conceptual and normative foundations. Biological Conservation, 2023. 288: p. 110350.

[4]. Gbedomon, R.C., et al., Quantifying the Diversity of Normative Positions in Conservation Sciences. Conservation, 2025. 5(3): p. 38.

[5]. Chevene, F., S. Doleadec, and D. Chessel, A fuzzy coding approach for the analysis of long-term ecological data. Freshwater biology, 1994. 31(3): p. 295-309.

[6]. Dray, S. and A.-B. Dufour, The ade4 package: implementing the duality diagram for ecologists. Journal of statistical software, 2007. 22: p. 1-20.

[7]. Sideropoulos, C. and A.Y. Troumbis, Conservation Culturomics 2.0 (?): Information Entropy, Big Data, and Global Public Awareness in the Anthropocene Narrative Issues. Earth, 2025. 6(2): p. 22.