UNIVERSITY OF CALIFORNIA,
IRVINE

Discovering latent social concepts across diverse societies

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical Behavioral Sciences

by

Maryam Gooyabadi

Dissertation Committee:
Professor Louis Narens, Chair
Professor Jean-Paul Carvalho, Chair
Professor Zyg Pizlo

2021

DEDICATION

To daemons chopping wood and carrying water
and to the ones who make the journey to Ithaka marvelous

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to the wondrous individuals who have made this work possible.

My heartfelt appreciation to the faculty members who supported me as a person and a scholar during my graduate career at the Institute for Mathematical Behavioral Sciences. To the savant Louis Narens, my mentor and friend, through engaging talks and shared meals, walks in the park and ice cream bets, movie nights and book clubs, you are a constant source of joy and growth. To Jean-Paul Carvalho whose graciousness, unwavering enthusiasm, and acumen is a true inspiration. To Kimberly Jameson for her invaluable guidance and support. To Don Saari with his kind words of encouragement, always an uplifting presence. Much thanks to my other committee members, Erik Sudderth for his algorithmic recommendations and guidance, to Zyg Pizlo's feedback and thoughts, and to Michael Tesler for insights on political ideology. To my many wonderful RAs who participated in my lab over the years.

My warmest gratitude to the singular Nikhil, the perfect mate, for his character, ingenuity, insights, wisdom, and partnership. Methods in two of the chapters in this dissertation is in collaboration with Nikhil – to be co-published in the near future. Our collaboration on the SUREN project has been and continues to be profoundly enriching and fun. I look forward to our lasting companionship and future expansions.

To my mom, Hadi, Reza & Zaynab for their loving support. To the Adzuara clan for their kindness and generosity over the years. To Jayanti & Steve and all the Addlemans and Ghoshs for their warm welcome and care. To Siva, for all things big and small, across countless life events, your energetic brightness a constant source of love and solace. To Mar Mar, a lifelong friend who helped fill every summer with adventure. To Vlad, Hapu, & the tiny, who dearly saw me through this journey before it began till my defense.

To Mark, Kyle, and Arpita, my PhD gang from the first day till almost the last (thank you COVID). Our Friday night dinners and grocery trips, morning spicy Mama noodles with dumplings, the many New Year feasts (Lao, Nepali, Persian, Chinese, Diwali) provided rhythm to my weeks. To Ehsan, a dear neighbor and true friend who made Irvine feel like home. To Aria, Shantanu & Hillary, and Sidra, UCI wouldn't be the same without you.

To my IMBS friends, Kirbi, Calvin, and Lucila, although I forget the reasons for "*why we can't have nice things* " (I think I have an idea), our time together provided me with many nice memories. I especially want to thank Kirbi Joe, my co-author, collaborator, the other pea in the pod, and Snuggie enthusiast who made our work on projects magical. We joined IMBS together and are left together. I'm proud of us – John Sommerhauser would be too.

I am grateful for financial support from IMBS and MDP programs. Two chapters of this dissertation are reprints of material from the *Journal of the Optical Society of America A (JOSA)* and *Journal of Applied Mathematics and Computation of Elsevier (AMC).* The JOSA publication was funded by a UC Pacific Rim Research award, 2010-2015 (Jameson, PI) and a research grant from the NSF 2014-2018 (SMA-1416907, K.A. Jameson, PI).

CURRICULUM VITAE

Maryam Gooyabadi

EDUCATION

Doctor of Philosophy in Mathematical Behavioral Sciences **2021**
Masters of Art in Philosophy, Political Science, Economics **2018**

University of California, Irvine *Irvine, CA*

Bachelor of Arts in Psychology **2015**
City University of New York *New York, NY*

Bachelor of Science in Computer Science Engineering **2011**
University of Colorado, Boulder *Boulder, CO*

REFEREED JOURNAL PUBLICATIONS

Joe, K. **Gooyabadi, M.**, (2021). A Bayesian nonparametric mixture model for studying universal patterns in color naming. Applied Mathematics and Computation, Elsevier, vol. 395(C).

Gooyabadi, M., Joe, K., & Narens, L. (2019). Further evolution of natural categorization systems: an approach to evolving color concepts. *JOSA A*, 36(2), 159–172.

WORKSHOPS

Graduate Workshop in Computational Social Science II **June 2021**
Santa Fe Institute

Internet-based Data Collection and Analysis in Judgment and Decision Making **September 2019**
Universität Konstanz

GRANTS, FELLOWSHIPS, HONORS, AND AWARDS

DTIE Summer Graduate fellowship – Online course design, UC Irvine
Multidisciplinary Design Program, UROP, UC Irvine
Christian Werner Fellowship, UC Irvine
Associate Dean's Fellowship, UC Irvine

ABSTRACT OF THE DISSERTATION

Discovering latent social concepts across diverse societies

By

Maryam Gooyabadi

Doctor of Philosophy in Mathematical Behavioral Sciences

University of California, Irvine, 2021

Professor Louis Narens, Chair

This dissertation is a multidisciplinary approach that brings together computational methods in machine learning to aid quantitative methods in the social sciences towards the study of **social conventions** - in particular, linguistic meaning and ideologies. Recent advancements in machine learning have yet to be applied in the social sciences where they can help identify groups with distinct underlying properties in order to gain insight into their unique conventions. Here, *social conventions* can be thought of as regularities of behavior (eg. norms) and beliefs that are shared between members of a group, those that govern social interactions or ascribe meaning to actions which form through tacit agreement. We used universal features common to all groups as a means of identifying latent groups present in the data. This approach reveals extant patterns without relying on prior assumptions, cultural knowledge, or predefined subgroups to highlight endogenous features within and between groups. The four projects study various aspects of social conventions, identify salient concepts important to groups, and model mechanisms that drive group beliefs and behavior. Much of these studies are dedicated to developing and testing effective methodologies and findings from each study informs aspects of consequent ones. These studies consist of: 1. Universal schema of the World Color Survey (WCS), 2. Cultural Consensus of Ideological groups, 3. Group probabilistic ordering of moral concepts, and 4. Further evolution of natural categorization systems.

Chapter 1

Introduction

This dissertation develops novel computational and mathematical approaches in the study of social conventions. Social conventions can be thought of as regularities of behavior and beliefs that form through tacit agreement and are observed by members of the group. Namely, the linguistic convention of color naming serves as a base case to create and test new techniques before extending such methods to the study of a more difficult social convention, ideologies or system of shared moral beliefs. The central task of this work is a methodological one: How to best analyze said social conventions with minimal researcher and tool biases. Here, we will utilize techniques in Machine Learning (ML).

Over the course of this work various aspects of these conventions will be modeled. In Chapter 2, unsupervised machine learning is deployed to identify universal or "natural" groups based on patterns of color naming across the 2,500 participants in 104 language groups in the world color survey (WCS) [68]. The methods are then extended to study ideologies in Chapter 3. Chapter 4 proposes a chess scaling algorithm to study rank order data of moral concepts. Chapter 5 uses reinforcement learning to study salience of concepts across participants of the WCS. The last chapter concludes with considerations informed by these studies for modeling the formation and evolution of social conventions, specifically, ideologies.

1.1 Background

Lewis formalized the *conventionalists* stance, as first presented by Hermogenes' (see 1.1.1), in his 1969 book *Convention.* Consistent with this formalization, in this paper, a convention can be thought of as regularities observed by members of a group that are invented, artificial, or optional. It is important to note that not all regularities are conventions, for example, neither eating nor sleeping or any other natural or fundamental human activities are. The manner in which we do so however, the customs of dress or cuisine is. That is to say that these conventions are a matter of choice that emerge through social interactions. Two of the earliest and most commonly debated conventions are language and money. Later, Lewis added beliefs to the list of conventions as well ([75]). Language and beliefs are of central interest in this dissertation work.

1.1.1 Conventions

Arguments about what is a convention, how it arises, and which phenomena fall under such a category has a long history in philosophy. Among philosophers today the term "convention" most closely follows the 4th century Athenian philosopher, Hermogenes' conception. as written in *Cratylus*([102]), defending a conventionalist view of language, Hermogenes says this of linguistic meaning:

> [N]o one is able to persuade me that the correctness of names is determined by anything besides convention... No name belongs to a particular thing by nature, but only because of the rules and usages of those who establish the usage and

> call it by that name, (384c-d)

A statement starkly opposed to Cratylus' anti-conventionist, or "naturalist" stance :

> A thing's name isn't whatever people agree to call it —some bit of their native language that applies to it—but there is a natural correctness of names, which is the same for everyone, Greek or foreigner. (383a-b)

Here, we see the stark difference between the *conventionalist* and the *naturalist* schools of thought: names or words are chosen arbitrarily to refer to certain objects versus names naturally *belong* to these objects.

Further Hermogenes thought that the mechanism that gives rise to conventions are through tacit social interactions. Hermogenes' central idea is echoed in Aristotle's De Interpretatione [10] and later, described by David Hume as arising *"gradually ...[and] without any explicit promise."* is shared amongst most philosophers today [54]. Convention as an implicit agreement contrasts with earlier ideas of conventions forming though explicit covenant. These stated promises or *social contracts*, as Thomas Hobbes's theory of government goes [44], takes agents out of the horrid *state of nature* and into more suitable conditions. While some conventions may indeed result from binding treaties, most seem to arise in the absence of such formalities or convening. Particularly, John Locke [80] underscores the importance of *tacit* agreements where individuals behave *as if* there is a prearranged set of rules that dictate behavior, while never formally expressing such rules. The notion of tacit agreements while insufficient to explain language as a convention, none the less, established the idea of convention as a set of implicit processes. While Bertrand Russell and Quine, amongst others, discarded the appeal to these *as if* agreements, David Lewis [76] was the first to propose a systematic theory of how social convention produces linguistic meaning. Through *signaling games*, Lewis shows how linguistic conventions form. Specifically, a *communicator (C)* possesses information about some *state of nature (p)* for which they send a signal to *audience (A)* who responds with an action. The need to coordinate between A and C where C's signal must elicit a mutually desired action from A results in a convention forming between A and C. In later works, Lewis amends his theory of convention as a *"regularity of action"* to also include beliefs as convention. When C signals to a multitude of A's, they can select the one that's actions are corresponding to their desired outcome. This A is someone who believes p, or believes that A believes p. Hence, a link is established between language and beliefs. In *"Languages and Language"* [75] the definition of convention was broadened to regularities of action *and belief.*

Earlier, Hume also established the idea that beliefs are an integral part of conventions. In the *Enquiry Concerning Human Understanding*[55], he defines a convention as:

> a sense of common interest; which sense each man feels in his own breast, which he remarks in his fellows, and which carries him, in concurrence with others into a general plan or system of actions, which tends to public utility. (p. 257)

This definition shares common features with Lewis' later conception: there is mutual benefit that arises from shared conventions, it arises through tacit agreements, each person believes the other person follows the convention, such beliefs results in following the convention themselves. Thought this way, social order can arise from individual coordination with group conventions and it need not form through explicit agreement or centralized authority.

In this work, ideology is thought of a system of shared beliefs, a social convention that sanctions ideas, behavior, and norms. Similar to linguistic meaning, ideology allows for mass coordination across group members and are learned through social interactions or observation. There is a long record of studying language as a convention with a sizeable community of active scholars. Ideology on the other hand has little to none. That is not to overlook the age long philosophical debate on ideology but rather to emphasize the need for a combined scientific effort towards the study of ideology. The following sections provides background on the linguistic conventions on color naming and a succinct, millennial's worth summary on ideology.

1.1.2 Linguistic convention - Color naming

Each person's experience of color depends upon the number of cones in the eyes, how their brain interprets the wavelengths reflected off objects, and the level of importance their culture assigns to hues. While the human eye can detect millions of colors, the small discrete set of terms divides the continuous color space into regions: “Red”, “Blue”, etc. Color naming has a long, rich history of scholarship dating back to Gladstone's *Studies on Homer and the Homeric Age:*(1858) [37] analysis of Homer's use of color words. Rather than finding the ocean blue in the Odyssey and the Iliad, Gladstone found peculiar color naming patterns: “the wine dark sea”, “rams were violet wool” and “iron”, “honey” and “faces torn with fear” were green. The most frequent color was “Black” and “White,” no blue in sight. He then came the controversial conclusion that the ancient Greek must have been color deficient.

Gladstone however was not alone in his discovery of odd color naming habits and the lack of “blue” in ancient lexicon. Philologist Lazarus Geiger (1870) [35] also noted the distinct color naming patterns Icelandic sagas, Hindu Vedic hymns, the Koran, ancient Chinese stories, and an ancient Hebrew version of the Bible. For the Hindu Vedic hymns he simply stated:

> These hymns, of more than ten thousand lines, are brimming with descriptions of the heavens. Scarcely any subject is evoked more frequently. The sun and reddening dawn's play of color, day and night, cloud and lightning, the air and ether, all these are unfolded before us, again and again ... but there is one thing no one would ever learn from these ancient songs ... and that is that the sky is **blue**.

It was not that blue did not exist in any ancient culture. In fact, about 6,000 years ago, humans discovered a naturally occurring blue, Lapis, a semiprecious stone mined in Afghanistan. Egyptians were the first to name the color blue and attempt at developing blue paint [114]. When combines with solutions, the Lapis turned dull and grey so instead it was used to make jewelry, headdresses, and applied to pottery (see 1.1). Through trade, royalty in Persia, Mesopotamia, and Rome gained access to Lapis and adorned their regalia. Due to its price, blue was exclusively used by royalty until it was manufactured and used by the church. At 431 AD, the Catholic Church assigned blue to Virgin Mary's robe (see 1.2) making it a symbol of innocence and trust worthiness. Borrowing from such aura, the military and police also wore blue and change it into a color of authority. As later shades of blue were developed, new names for blue entered the language.

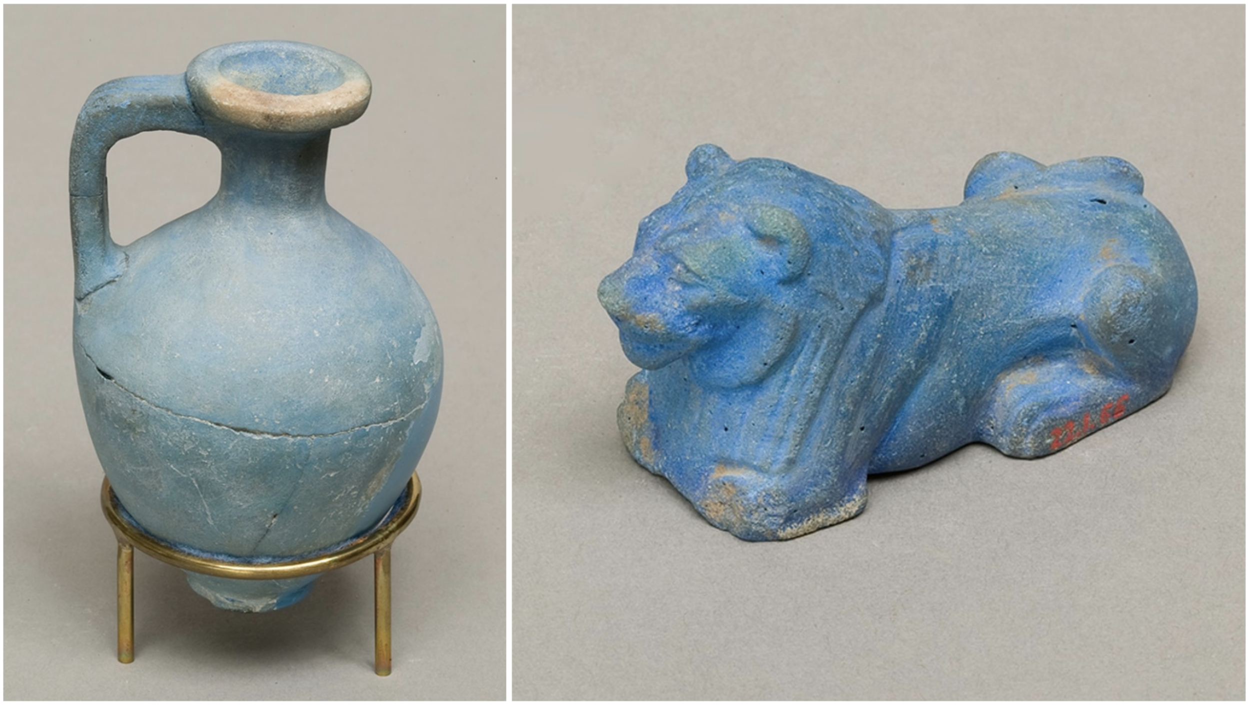


Figure 1.1: Left: Egyptian Juglet, ca. 1750–1640 B.C. (Photo: Met Museum, Rogers Fund and Edward S. Harkness Gift, 1922. (CC0 1.0)). Right: Figure of a Lion. ca. 1981–1640 B.C. (Photo: Met Museum, Rogers Fund and Edward S. Harkness Gift, 1922. (CC0 1.0)) [127]

Both William Gladstone and Lazarus Geiger had proposed a universal evolutionary sequence color vocabulary. While they mistakenly assumed vocabulary evolved in tandem with the

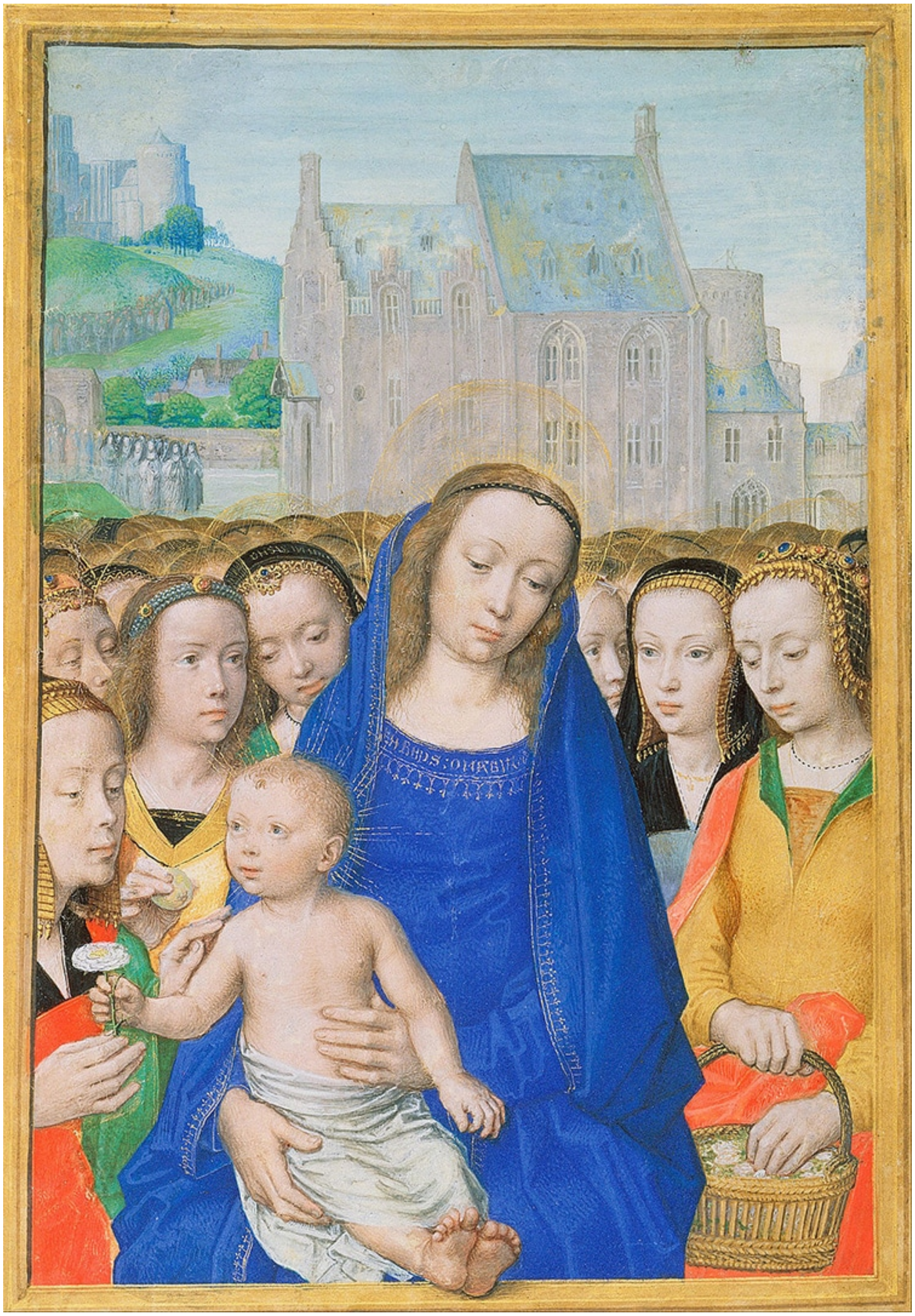


Figure 1.2: “Virgin and Child with Female Saints” by Gérard David, 1500. [95]

evolution of biological color sense – rather than the industrial production ability of color – their predictions were found support in later studies by Brent Berlin and Paul Kay. Belin and Kay developed the notion of Basic color terms and studied the cultural variation in assigning names to hues. While there are little to none variability in these language groups' visual ability, the number of basic color terms varies greatly (i.e. between 3 to 12). The Himba tribe of Namibia is one such example [109] where the lack of word for "Blue" results in participants' inability to identify the blue square in the right side of figure 1.3, labeling all the hues as *Buru*. Yet with "Green" as culturally important, they distinguish the green hue circled on the right part of the figure as *Dambu* from the rest.

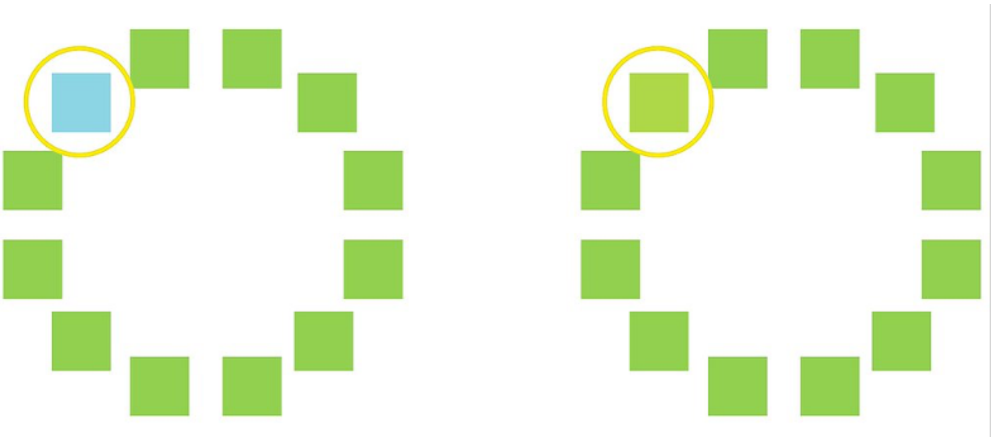


Figure 1.3: Left: all hue tiles are labeled with the same word *Buru*. Right: the tile circled is assigned the name *Dambu* while the rest *Buru*

In the the following sections the summary of work on color naming by Berlin and Kay is presented.

(i) Basic Color Terms

When Berlin and Kay published *Basic Color Terms* [16], they revived the universality view in the color naming debate by drawing two primary conclusions about the acquisition and evolution of color terms: (i) that there was a limited set of color words from which all languages derived their color terms and (ii) that languages acquired these terms in a moderately fixed order. These terms were coined by Berlin and Kay as *basic color terms*—the smallest set of color category names with which a person could name the entire color space. Basic color terms were determined according to the following semantic and syntactic criteria:

1. It is monolexemic (blue, not blue-green)
2. It is monomorphemic (blue, not bluish)
3. It is not contained in another, broader color category (scarlet is a type of red)
4. Its use is not limited to a narrow class of objects (blond is usually restricted to hair and beer colors)
5. It has to be psychologically salient to the general population ("the color of my car" is not salient to most people).

There are 11 basic color terms identified for English: white, black, red, yellow, green, blue, orange, brown, purple, pink, and gray. These 11 terms also serve as the proposed limited set from which all languages draw their color terms. Based on these criteria, Berlin and Kay concluded most languages to have between 2 and 11 basic color terms.

Additionally, Berlin and Kay found that languages followed a fixed evolutionary pattern that they organized into a seven stage evolutionary path. They posited that languages acquired new color terms in the following order:

1. Stage I: Dark/cool (black) and light/warm (white)
2. Stage II: Red
3. Stage III: Either green or yellow
4. Stage IV: Both green and yellow

5. Stage V: Blue

6. Stage VI: Brown

7. Stage VII: Brown, purple, pink, and gray

Several years after *Basic Color Terms* was published and in response to new research on this topic, this evolutionary hierarchy was amended into the five stage hierarchy depicted in Figure 1.4. This amended hierarchy still proposed that languages acquired color terms in a fixed order but now presented several possible paths by which languages could incorporate these new terms into their existing system. In summary, the seminal work of Berlin and Kay provided a platform for view that color naming is driven by universal principles, not relative, culture-specific phenomenon, and reignited the long-standing debate regarding the origins of color naming conventions.

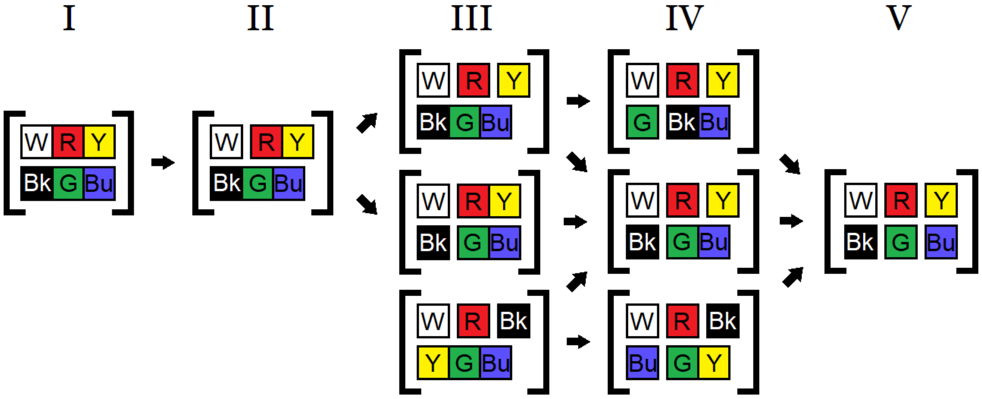


Figure 1.4: Evolutionary stages and their corresponding basic color term (BCT) systems, using English term equivalents. Stages are determined by the number of BCTs in the system. Colored blocks that are connected represent a single term which encompasses all of those color regions. For example, Stage I is a 2 BCT stage where one term refers to the white/red/yellow region of the color space and the other term refers to the black/green/blue region. All 110 languages from the WCS were assigned to either one of the 9 systems depicted or to a transitional state between two consecutive stages.

(ii) World Color Survey

In 1969, Brent Berlin and Paul Kay published their book *Basic Color Terms: Their Universality and Evolution*, which sparked a renewed interest in the study of color naming in academia. The WCS data was collected from 110 unwritten, monolingual, pre-industrial, tribal languages, with an average of 24 participants per language (∼2,640 participants in total). Participants completed two tasks: the *naming task* and the *mapping task*. In the naming task, participants assigned names to 330 Munsell color chips (see Figure 1.5, which were presented one at a time, in a fixed random order. In the mapping task, participants were given a color term from their language and were asked to pick a color chip (or set of color chips) from the set of 330 which best exemplified that term. This set of color chips are called *focal colors*. The data is publicly available at no cost via the project website (http://www1.icsi.berkeley.edu/wcs/data.html). The data used in our methods is the mapping task data (i.e. term.txt file on the project website). The data text sheet consists of a long list of the names participants from all languages assigned to each color chip where each line contains the language number (1 to 110), participant number, chip number (1 to 330), and abbreviation of term used. Supplementary text files contain information on each language, participant details, as well as data from the mapping task.

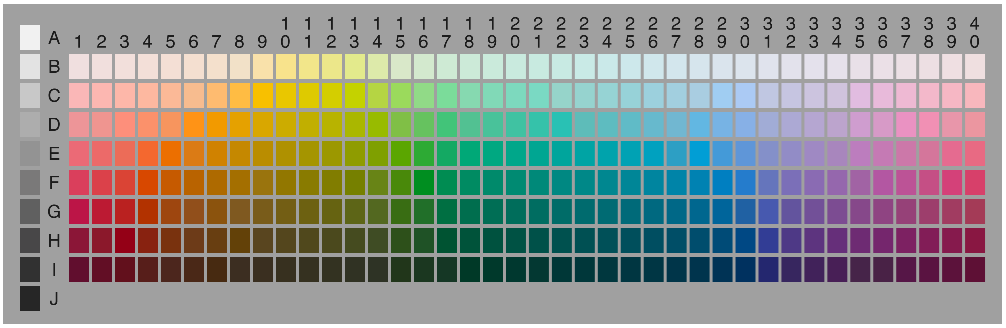


Figure 1.5: The set of 330 Munsell color chips used in the World Color Survey.

Included in the material that was released in their book [16] was an empirical study, seemingly supporting Berlin and Kay's theoretical claims, which included color naming data for a small sample of speakers. This study was met with numerous criticisms. The critiques of the empirical study were that (i) the sample was too small (20 total participants with three or fewer speakers per language), (ii) the speakers were bilingual (spoke their native tongue plus English), (iii) the data was collected in Northern California instead of their native communities, and (iv) the languages represented by the sample were mostly from industrialized societies [67]. In response to these critics, Berlin and Kay embarked on a project called the World Color Survey. They addressed the critiques of their previous empirical work by collecting a much larger sample consisting of monolingual speakers from pre-industrial, tribal languages worldwide. The final data set was comprised of 110 languages, each with 24 participants on average (modal number was 25) [22].

The World Color Survey was given to participants in two tasks: the *mapping task* and the *naming task*. The *mapping task* was a focal identification procedure. The facilitator would present a word to the participant (a list of basic color terms would have been first elicited from the participant) and then the participant would indicate which color chip in the Munsell grid was the best exemplar of that word. The *naming task* instead presented participants with a color chip and prompted them for a name for that color. For this task, participants were asked to name all 330 color chips in Figure 1.5 one at a time in a fixed, random order. Once compiled, the *naming task* data for a participant would create a full partition of the color space. Kay and colleagues [68] concluded that the findings from the World Color Survey were in agreement with the original claims made by Berlin and Kay in *Basic Color Terms*.

1.1.3 Ideology

Ideology is and has been a central topic within the social sciences and philosophy, yet the literature reveals varied and often opposing definitions of ideology and refers to disparate social phenomenon that may be related in a larger process –eg. psychological needs of individuals drive ideologies versus ideologies being imposed by institutional powers and society. Yet, the causal mechanisms through which it shapes social and political behavior remain poorly understood. Importantly to this study, the term ideology is not relegated only to the political sphere (eg. voting outcomes or party agenda) but rather as a broader concept that can include any shared system of beliefs. This can include religious and cultural ideas that sanction particular thoughts and behavior –both on the personal and social level. In the following studies, any group may have an ideology.

(i) History of the concept of ideology

The word ideology was not coined until 1801 but the early conceptions of ideology go as far back as Aristotle. A common conception related to the discrepancy between objective reality and the social conceptions or interpretations of these realities –i.e. the social narratives around these realities– with no causal mechanism as to how such realities arise. Such ideas had been noted by Marsilio of Padua and by Machiavelli in their Discourses on Livy [118]. Similarly, Bacon studied the preconceptions and "illusions" of the populace (*praenotiones vulgares*). He spoke of the tendency of the mind to take ideas of things as the things themselves, and then develop knowledge around these *idols*[21]. Such mental tendencies hindered the development of scientific knowledge and enlightenment. These ideas pointed towards the irrationality of human more than a social convention.

It was not until the 17th century that ideology was thought of in a more socially constructed way were "climates of opinion" became synonymous with "instinctively held preconceptions," "conventional world-views," "basic intellectual viewpoint," "spirit of an age" (Zeitgeist), "Weltanshaunung", "intellectual climate," "collective 'state of mind,'" "the prevailing psychological state," and "national mood." Mainly, here ideology is the generality of opinions, specific to a historical era present even in the historians' accounts of "historical facts" [115]. In Glanvill's words, *'the larger Souls, that have travail'd the divers Climates of Opinions'* [38], and who are 'more cautious in their resolves, and more sparing to determine' elevate their ideas to be the same as commonsense. Not only was ideology dependent on the times but humans' mentalities were conditioned by such systems.

However by the time De Tracy coins the term *ideology*[26], he was referring to the scientific study of ideas and their origins. Rather than ideology arising through natural mental tendencies, they were the result of material realities and forces that shape people's thoughts. As a proponent for social progress through political policies, De Tracy believed ideology would enable efforts on behalf of human progress. Falling out of favor with Nepolean during his political campaigns, the later labeled De Tray and his followers as 'ideologues... [the] unrealistic idealistic fanatics... [and] mongers of metaphysical trash'[26]. This was the first time ideology was given a negative connotation and popularized as such, especially in reference to political ideologies.

This departure from the study of ideas towards viewing ideology as irrational, similar to earlier conceptions, grew with the conviction that human behavior is largely non-rational or "irrational" (eg. [91][90][134][9]). Darwin's work was a driving force behind such beliefs as he saw forces that shaped human behavior, as other species, no different from those influencing that of animals –i.e. largely instinctive or "non- logical." Therefore, in such conceptions, human social behavior was also driven by non-rational forces (eg. [56][89][123][128][129][130][52]).

Our modern day concept of ideology is greatly affected by Marx and Engels' theories on ideology formation[26], in large part made prominent under the Communist rule of Stalin and Mao. The Communist Manifesto [87], a commissioned work written by Marx, outlines the internal conflicts intrinsic within Capitalism and the class struggle modern workers experience– rather than the form Communism would (or should) take. In fact, Marx said little about Communism as a form of governance or economic system partly because his theory of social change emphasized an inevitable evolutionary process towards a different economic system. Communist leader took Marx's writings to be reductio ad impossibile or proof by contradiction for Communism's superiority over Capitalism and attempted to usher in this social change resulting in catastrophic failure and a dear loss of human life. It is not the this short commissioned pamphlet where Marx outlines ideology but in his later work, Capital [86], that ignited a long debate in the fields of sociology and later, political science. In Capital, he positions the material world of economic relations, particularly in the production process, as the central pillar that gave rise to ideology: ideology as a function of the material reality of production. Largely "unconscious" processes, through exploitative and alienating features of capitalist economic relations, workers and the ruling class alike develop an ideology that protects private property and maintain the legitimacy of the owners of production. Society's "productive forces" learned by workers coupled with "productive relations" or "property relations" create the economic "foundations" for which the political and legal "superstructure" with its "ideologies," including religion, art, science, philosophy, and morals are created. Marx and Engels presented an entire process that created and maintained ideology in society.

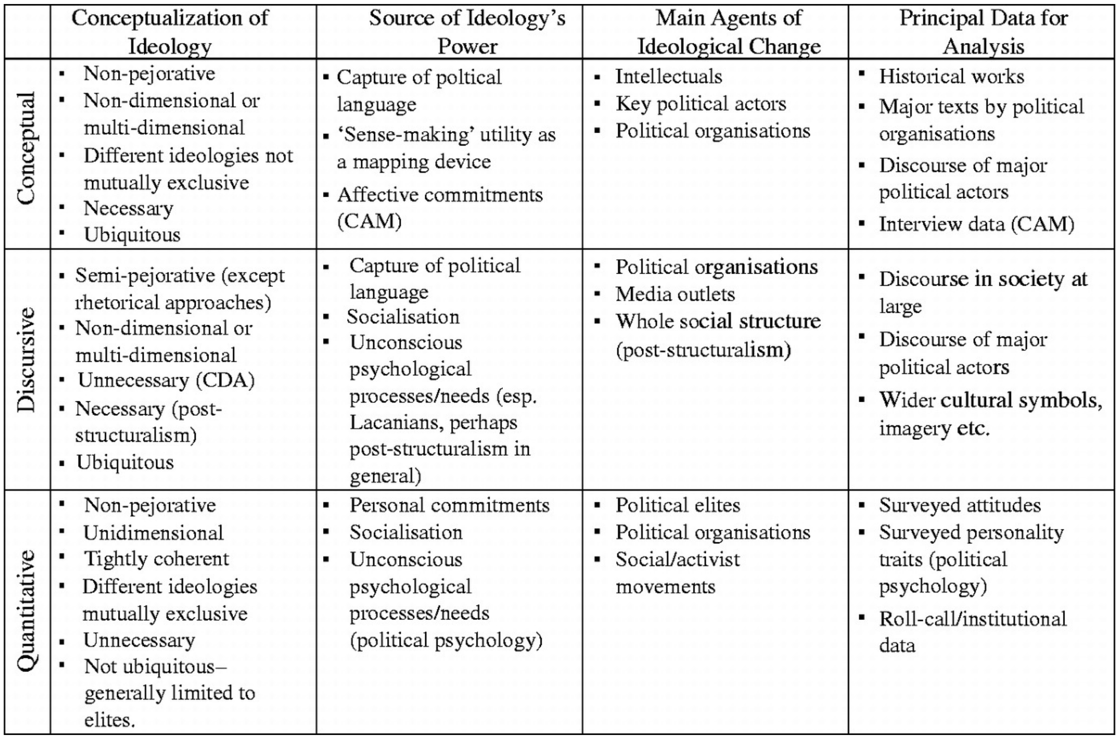
In response to Marx, with his "deterministic" depiction of ideology, Sorel, Durkheim, and Weber presented other causal models of ideology[25]. Sorel [122] while not using the term "ideology," instead spoke of "myths" as "a body of images capable of evoking sentiment instinctively," that were adopted by people in societies while Durkheim –similar to Bacon, aspired to eliminate ideological thinking from the social sciences– spoke of "doctrines" as a form of rationalization for preconceived ideas present in society. Most importantly, Weber in his seminal work, *Spirit of capitalism*[132], presents his Protestant Ethic thesis that analyses the promises of salvation and its role in capital accumulation in the Protestants. Rather than arising from class relations, Weber sees ideas in the realm of religion bringing about material circumstances that give rise to capitalistic societies. As for the ideology that is promoted by the politicians or ruling class, is his “rationalization thesis” presents ways in which politicians use ideology to legitimize and rationalize pursuing the end that they had already decided before. Thinkers who followed, such as Mannheim [85], saw ideology as an attempt to explain the course of history . Instead of an evaluative view on ideology, somewhat similar to De Tracy, he understands ideology as a relational process of constructing knowledge that is itself influenced by the possessor of knowledge, especially historians’ construction of knowledge. In the same vein, Cohen also views ideology as one of many possible explanations, endorsed by judicial decisions affected by political ideas of the time.

The prescriptive nature of ideology was endorsed by Daniel Bell who believed it to be “an action-oriented system of beliefs” that is not to make reality understandable but rather to motivate individuals towards or away from certain actions. As such, ideology is a process of justification that requires the obfuscation of reality. And finally, this brings up to 1971 where Althusser [2], amongst all the other prominent Noe-Marxists –such as Lukács, Gramsci, Adorno, Horkheimer, Debord– presented his influential analysis of ideological state apparatuses in a macro system where culture is the central unit of analysis independent from “mode of production”.

These thinkers fall into different schools of thought yet, are often responding to each other’s work. This contrasts greatly with how ideology is spoken about today: fragmented into various departments with little discourse to bridge this gap.

1.1.4 (ii) Approaches in the study of ideology

The collective works of the early thinkers mentioned in the past section emphasized processes through which ideologies shaped social behavior, in particular in the political realm and served as a stabilizing force for extant power structures ([33][34][85][92]). The extent to which those is power forced ideologies on masses was a matter of heated debated. Yet, since the mid to late 20th century, the study of ideology and the discourse around ideology has progressively splintered, generating substantial methodological and theoretical cleavages. Studying these isolated approached to ideology, Leader Maynard [88] noted three distinctive methodological orientation in the existing scholarship of ideology: conceptual, discursive, and quantitative (see figure 1.6). Where conceptual approaches analyze component concepts and the content of distinct ideologies. The discursive approaches, as the name would suggest focus on text, speech, and non-verbal discourse such as imagery and symbols (i.e. the expression of ideology) and its role in shaping political beliefs and behavior. Finally, the quantitative approaches focus on discovering underlying relationships between individual traits and political attitudes, using survey and experimental data.



	Conceptualization of Ideology	Source of Ideology's Power	Main Agents of Ideological Change	Principal Data for Analysis
Conceptual	▪ Non-pejorative ▪ Non-dimensional or multi-dimensional ▪ Different ideologies not mutually exclusive ▪ Necessary ▪ Ubiquitous	▪ Capture of poltical language ▪ 'Sense-making' utility as a mapping device ▪ Affective commitments (CAM)	▪ Intellectuals ▪ Key political actors ▪ Political organisations	▪ Historical works ▪ Major texts by political organisations ▪ Discourse of major political actors ▪ Interview data (CAM)
Discursive	▪ Semi-pejorative (except rhetorical approaches) ▪ Non-dimensional or multi-dimensional ▪ Unnecessary (CDA) ▪ Necessary (post-structuralism) ▪ Ubiquitous	▪ Capture of political language ▪ Socialisation ▪ Unconscious psychological processes/needs (esp. Lacanians, perhaps post-structuralism in general)	▪ Political organisations ▪ Media outlets ▪ Whole social structure (post-structuralism)	▪ Discourse in society at large ▪ Discourse of major political actors ▪ Wider cultural symbols, imagery etc.
Quantitative	▪ Non-pejorative ▪ Unidimensional ▪ Tightly coherent ▪ Different ideologies mutually exclusive ▪ Unnecessary ▪ Not ubiquitous–generally limited to elites.	▪ Personal commitments ▪ Socialisation ▪ Unconscious psychological processes/needs (political psychology)	▪ Political elites ▪ Political organisations ▪ Social/activist movements	▪ Surveyed attitudes ▪ Surveyed personality traits (political psychology) ▪ Roll-call/institutional data

Figure 1.6: Descriptions of theoretical fields in the study of political ideology.[88]

These approaches differ in their unit of analysis, conceptualization of ideology, and methodology, which leads to two main incompatible ways to view ideology: 1. As micro systems of relationships between individuals and their attributes (eg. physiological and genetic predispositions [100][46][4][20][65][112] or as a macro system with social and institutional factors (eg. [98][124][32][**?**][27]); 2. ideology as spatial concepts with possible independent dimensions (eg. Extreme right conservative to Extreme left liberal; [103][42][113]) or non-spatial with an emphasis on ideologies as systems of beliefs rather than opposites (eg.[34][98][31]). There are attempts to integrate micro-macro processes of ideology ([42][64][43]) yet even in these attempts the social institutional forces remain marginal. A complexity approach to the study of ideology [51] attempted to outline underlying mechanism of ideology and possible evolution of ideology as a "system using as conceptual networks of cognitive-affective representations embedded in social networks of people" . In this approach, ideology groups have positive and negative associations with concepts, cognitive-affective maps, that are engaged like neurons when confronted with group salient concepts (eg. Government spending, illegal immigration). Yet, even in this approch, the semantic space created for each group is independent from within group consensus, learning, and interactions as well as between group competitions.

1.2 Conclusion

This chapter serves as an introduction to the concepts central in the ensuing chapters and serves as a rationale for the approaches undertaken. There is a long philosophical history in the conception of social conventions and separately so, ideology. As mentioned earlier, language as a social convention is one of two most prominent examples of conventions and there is a great scientific community that have studies its many fundamental aspect. Ideology on the other hand has not received the same treatment: philosophical debates about what it is or its underlying processes remain theoretical while most empirical approaches are disjointed, studying and labeling many disparate phenomena under the umbrella term of ideology. This work, as presented in the following chapters, attempts to study ideology similarly to how linguistic convention is studied. By outlining methodology to study a simple case of language (i.e. color naming) and extending such methodology to ideology, this dissertation serves as a starting point to the scientific study of ideology. The challenges we face when doing so is three fold: first, ideology is a complex phenomenon that has visible and non observable elements; Second, There are no standard data sets that grounds research efforts from a multitude of fields as it is in color naming; Third, as a consequence of the second challenge, there is not a community of scholars whose efforts build upon each other to provide insights relevant across domains. Let us explore how to approach ideology with these three challenges in mind.

Chapter 2

A Bayesian nonparametric mixture model for studying universal patterns in color naming

Variational Inference for the Beta-Bernoulli Dirichlet Process Mixture Model is employed to uncover universal patterns in color naming systems. The data used consist of 2,552 participants from 106 World Color Survey languages. To study these languages collectively, the model is informed by universal biological, linguistic, and topological features of the task. We find that the majority of the naming systems are represented by eighteen clusters, each constituting a universal pattern. Novel mathematical techniques are developed to study the levels of similarity, underlying consensus, and diversity among these patterns. This implementation of nonparametric models demonstrates how machine learning methods can be tailored for behavioral science applications.

2.1 Introduction

This paper is aimed at two groups: social scientists interested in utilizing Machine Learning (ML) techniques and computer scientists applying ML algorithms to social science topics.

Scientific inquiry involves understanding universal properties or laws that drive natural or social phenomena. The two approaches to studying universality include (i) taking micro instances as the unit of analysis to generalize to macro phenomena (e.g. generalizing insights from English speakers to all languages) and (ii) starting from macro patterns to explain instances in micro phenomena. Methodologically, researchers in comparative studies usually use the first approach. Yet, with recent advancements in computation power and ML techniques, the second approach to universality is now more realistic. Researchers can base their ML models on shared universal features (e.g. physiological features of vision) rather than a myriad of factors (e.g. age, gender). This kind of approach reveals extant patterns without relying on prior assumptions, cultural knowledge, or predefined subgroups. Instead, it lets the data identify the number of "natural" groups present and highlights endogenous features within and between these groups.

However, advanced ML techniques can be general in nature, requiring specific customization before used effectively for particular social phenomena. Such customization requires intimate knowledge of the social phenomenon and the ML algorithms, as well as the programming skills for implementation. Hence, there is great need and opportunity for collaboration between computer and social scientists on the application of ML techniques to social science topics. ML has great potential to revolutionize social scientific inquiry and open doors to new and important discoveries, not dissimilar to its impact on predictive algorithms used in online commerce or media.

This paper demonstrates the work of both sides through a specific example: it uses the second approach to universality by customizing advanced ML techniques based on key features of color naming systems. We investigate patterns in color naming and take the following universal properties as the impetus of the model:

1. Biological and physiological features on the human eye results in color vision;
2. The need to communicate about color is a universal linguistic feature and to do so requires categorizing colors into linguistic terms;
3. Color is a continuous space with its own unique topology.

Color is a particularly useful example given the universality of its concept and its well-established scholarship—spanning decades and many disciplines (see Section 2.1.1). The typical approach to color naming has been to draw inference about shared meaning across languages by examining meaning within each language. Conversely, we develop a methodology which characterizes the data without reference to language or culture and use unsupervised machine learning (i.e. Beta-Bernoulli Dirichlet Process Mixture Model with Variational Inference) to provide an efficient means for carrying out the second approach. This methodology can be extended to many other social science investigations where shared universal properties give rise to social phenomena (e.g. political group preferences, cultural ideas, cooperation). In all these applications, ML techniques can reveal latent patterns within the data, empowering researchers to derive new insights not feasible before.

Language is an integral part of society which enables communication among its members. To shed light on how words gain their meaning and how their meaning evolves over time, color naming is often used as a case study. The color domain can be defined by a physical space, making it a useful concept for studying denotation of meaning. Though humans can distinguish millions colors, language provides us with a small, manageable set of terms for categorizing the space. Partitions of the color space vary across different language groups and evolve over time (e.g. new color terms may enter a language). Investigating universal patterns in color naming provides insight into the mechanisms that give rise to the observed data. Recently, computational techniques have been utilized to study this phenomenon. Here, we develop a methodology for transforming a color naming data set— namely, the World Color Survey—which is based on constraints imposed by the stimulus space. This transformed data is used to initialize a nonparametric Bayesian machine learning model in order to implement a culture and theory-independent study of universal color naming patterns across different language groups. All of the methods described are executed by our Python software package called *ColorBBDP*.

2.1.1 The Study of Color Naming

A wide range of literature uses quantitative methods to understand the properties of linguistic categorization [120, 121, 117, 50, 23, 82, 19, 13], several of which have also been applied to the study of color naming. Color naming in particular has a long history in academia, beginning with seminal work on ancient Greek color terminology by Gladstone in 1858 [37] that was extended by other 19th century researchers to additional ancient languages. The subject gained increasing recognition in 1969 with Berlin and Kay's seminal work, *Basic Color Terms: Their Universality and Evolution* [16], and later from the *World Color Survey* (WCS) by Kay, Berlin, Maffi, Merrifeld, and Cook [68, 22]. Basic color terms, as defined by Berlin and Kay, are the smallest set of color category names with which a person could name the entire color space. This set is determined by evaluating each color term in a language against a series of linguistic criteria. The WCS was undertaken to validate the claims made in [16] and consequently produced a famous data set which includes color naming data for many languages worldwide. This data has been analyzed with a variety of methodologies including mathematical methods and ML techniques, such as simulations and clustering.

Among the mathematical methods applied to the WCS data, Fider et al. not only validated the findings of Kay and colleagues [29], but also drew further inference on features of the data not originally examined [28, 30]. In the realm of simulation-based models, Regier et al. [108] created optimal partitions of the 330 Munsell color chip set (see Figure 1.5) by maximizing a wellformedness measure, and provided quantitative evidence for Jameson and D'Andrade [57]. Following a series of literature by Jameson and Komarova [72, 71, 59, 60, 61, 58, 96, 101], Gooyabadi, Joe, and Narens [40] evaluated theories of color category evolution using WCS seed data in an agent-based model [39]. Additional agent-based simulations of color categorization conducted by Baronchelli, Puglisi, Loreto, et al. [104, 11, 81, 125] has yielded similar conclusions, further validating the use of these methods on this data.

In the clustering based ML methodology, Brown and Lindsey [77, 78] utilized k-means clustering algorithm on the WCS data. In [78], they clustered individual WCS participants' color naming systems from all languages to reveal universally occurring *motifs*—i.e. patterns of color term vocabulary and usage exhibited by a group of WCS participants. They found that (i) motifs were widespread and present in many unrelated languages, pointing to their universality, and (ii) there was a high level of diversity of motifs within languages.

Similarly, our model employs mathematical methodologies and unsupervised, clustering algorithms on the WCS data. The advantage of employing such quantitative methods is the ability to perform analyses independent of cultural and linguistic features of the language. Bayesian nonparametric models maintain this advantage and additionally eliminate computational assumptions in the modeling process.

2.1.2 Bayesian nonparametric models

In color naming there are no objective labels to assign to each color and participant systems are considered independent from their cultures, making this an unsupervised learning task. A key consideration to the model is that the true number of "natural" groups or clusters is unknown. Bayesian nonparametric models (BNP) assume an infinite number of latent clusters in contrast to the standard clustering methods which rely on a predefined, fixed number of clusters. The primary advantage of BNP is that the data determines the complexity of the model instead of complexity being ascertained via model selection *ex post facto* [36]. BNP uses a single, adaptive model that allows complexity to increase as more observations are introduced to the model. The model we employ is the Beta-Bernoulli Dirichlet Process Mixture Model.

Beta-Bernoulli Dirichlet Process Mixture Model with Variational Inference

The Beta-Bernoulli Dirichlet Process Mixture Model (BBDP) [97, 53] is a particular implementation of BNP. It combines a Beta-Bernoulli observation model with the Dirichlet Process mixture model. Together, it allows us to cluster binary features vectors of participants without assuming the number of clusters. It does so by using a Dirichlet process to estimate the number of Bernoulli distributions that likely generated the observed data. Model selection is performed by using variational inference methods [17] to estimate the lower bound of the marginal likelihood (i.e. the evidence of the lower bound or ELBO) of the observed data. The formalization of the model and its specifications are included in the following sections. Previously, this model has been used to successfully cluster discrete binary data, namely that of animal attributes, handwritten digits, and images of scenes [97].

2.2 Beta-Bernoulli Dirichlet Process Mixture Model with Variational Inference

The model used in this paper is called the Beta-Bernoulli Dirichlet Process Mixture Model using Variational Inference [97]. It is a specific case of the Dirichlet process mixture model which utilizes a beta-Bernoulli observation model—useful for data that is binary in nature—and variational Bayesian methods for inference. An in-depth description and formulation of the model is provided in the following appendix.

2.2.1 Beta-Bernoulli Observation and Mixture Models

Suppose we have data set $X = \{X_1, ..., X_N\}$, where each observation X_i is a binary vector with D dimensions representing D attributes of an observation. An entry $x_{id} = 1$ if X_i has the attribute d and $x_{id} = 0$ otherwise. If we let θ be the mean of the Bernoulli distribution, then the Bernoulli likelihood can be written generally as:

$$P(x|\theta) = \theta^x (1-\theta)^{1-x} \tag{2.1}$$

Using this form, the probability density for each observation X_i can then be computed by:

$$P(X_i|\theta) = \prod_{d=1}^{D} \theta_d^{x_{id}} (1-\theta_d)^{1-x_{id}} \tag{2.2}$$

where θ is a D-dimensional vector with entries θ_d for $d \in \{1, ..., D\}$ represent the probability that an observation has the attribute d.

The conjugate prior to the Bernoulli distribution is the Beta distribution with parameters β_1 and β_2. Therefore, the prior, $P(\theta)$ can be given by the following function:

$$P(\theta) = \frac{1}{\mathbf{B}(\beta_1, \beta_2)} \theta^{\beta_1 - 1} (1-\theta)^{\beta_2 - 1} \tag{2.3}$$

where the Beta function $\mathbf{B}(\beta_1, \beta_2)$ serves a normalization constant and β_1, β_2 are shape parameters that determined based on prior beliefs or existing knowledge. We search the only search a portion of the parameter space where $\beta_1, \beta_2 \in (0, 1)$ because the shape of the beta distribution is biased towards the bounds of its domain, 0 and 1, when $\beta_1, \beta_2 < 1$. This behavior is useful when drawing priors for the a Bernoulli mixture model.

A Beta-Bernoulli mixture model can be defined a mixture of K Beta-Bernoulli distributions. In order to identify which of the K distributions each data point X_i was drawn from, we introduce a latent variable $Z = \{Z_1, ..., Z_N\}$. For each $Z_i \in Z$, Z_i is a K-dimensional vector which has exactly one entry equal to 1, corresponding to the cluster assignment of X_i. Each of the K distributions in the mixture model has a corresponding weight, represented by $\pi = \{\pi_1, ..., \pi_K\}$, such that $\sum_{k=1}^{K} \pi_k = 1$. Therefore, the distribution of the latent variable Z conditioned upon its weights π is:

$$P(Z|\pi) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} \tag{2.4}$$

and the Bernoulli likelihood can then be formalized as:

$$P(X|Z, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} P(X_i|\theta_k)^{z_{ik}} \tag{2.5}$$

Figure 2.1 depicts a graphical model representation of the Beta-Bernoulli mixture model.

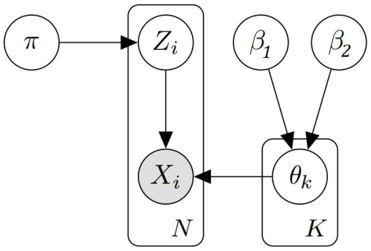


Figure 2.1: A graphical model of the Beta-Bernoulli mixture model.

2.2.2 Dirichlet Process Mixture Model

The Dirichlet Process (DP) is a nonparametric prior for infinite, discrete distributions. Therefore, the DP mixture model is able to cluster exchangeable data points without determining the number of clusters *a priori* by assuming an infinite number of latent clusters. For this reason, DP mixture models are synonymously known as infinite mixture models. These processes are commonly used in Bayesian nonparametric methods because they allow the number of clusters to grow as more data points are introduced to the model.

A DP can be thought of as a distribution over distributions. Suppose G is a Dirichlet process, then $G \sim DP(\alpha, G_0)$ where $\alpha \in R^+$ is called the dispersion parameter and G_0 is the base probability distribution. Draws from the process G are taken according to the following algorithm:

1. Assume there are $X_1, ..., X_N$ observations and k unique values for the variable K (which represent k clusters present at the time).

2. For observation X_i, with probability $\frac{\alpha}{N-1+\alpha}$, a new draw is taken from G_0 (i.e. X_i is assigned to a new cluster).

3. With probability $\frac{n_k}{N-1+\alpha}$, where n_k is the number of observations currently in cluster k, X_i joins cluster k.

4. Each observation is iteratively assigned to a cluster until all N observations have been grouped. Cluster assignments are stored in the latent variable Z where $Z_i \in Z$ is a K-dimensional vector with the k-th element being equal to 1 (corresponding to datum X_i being assigned to the cluster k) and all other elements equalling 0.

The end result of this process is then a distribution over the partitions of the data X, which serves as a prior over the class assignment vector Z. Some common analogies used to describe the DP are the Chinese Restaurant Process, the Stick-breaking Construction, and a modified version of Polya's Urn Scheme. A graphical model for the DP mixture model is presented in Figure 2.2.

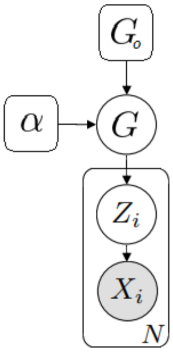


Figure 2.2: A graphical model of the Dirichlet Process mixture model.

2.2.3 Variational Inference

Due to the complexity of the statistical models in Bayesian nonparametric methods, many of the resulting integrals become intractible and thus require other techniques to approximate the parameters of the model. One such family of techniques is called *variational Bayesian inference*. Variational inference can be used as a way to (i) estimate the model's posterior distribution or (ii) to compute an evidence of the lower bound (ELBO), which is then used for model selection. The intuition behind (ii) is that the higher the computed marginal likelihood of a model is, the higher the probability that the data was generated by that model. Therefore, the model with the highest ELBO is selected as the most appropriate model, given the data. In this paper, we use variational inference for the purpose of computing the lower bound of the marginal likelihood.

Given a set of unobserved variables Z and a data set X, the posterior distribution can be approximated by the variational distribution Q:

$$P(Z|X) \approx Q(Z)$$

The aim of variational inference is to minimize the distance between the true posterior $P(Z|X)$ and the approximated distribution $Q(Z)$ and thus seeks to find the Q which minimizes this distance. The distance between distributions P and Q is most often formalized using Kullback-Leibler Divergence (KL-divergence), defined as

$$KL(Q \parallel P) = \sum_{Z} Q(Z) \log \frac{Q(Z)}{P(Z|X)} \tag{2.6}$$

This function can be rewritten and rearranged to yield

$$\begin{aligned} \log P(X) &= KL(Q \parallel P) - E_Z \left[\log Q(Z) - \log P(Z, X)\right] && (2.7) \\ &= KL(Q \parallel P) + \mathcal{L}(Q) && (2.8) \end{aligned}$$

The term $\mathcal{L}(Q)$ is called the Evidence Lower Bound (ELBO). Maximizing $\mathcal{L}(Q)$ will reveal the Q which minimizes the KL-divergence since $\log P(X)$ is fixed with respect to Q.

2.3 Methods

Color naming is an inherently linguistic task as it involves assigning a name (lexical term) to a color (stimulus). To discover universal patterns in color naming using only universal features, we must abstract away from each language's color terms. The universal features influencing color naming include biology, language, and the topology of the stimulus space. Therefore, our ML model is tailored to accommodate these universal features.

The color domain is a continuous space on which the human eye can discriminate millions of colors. To communicate easily about color, however, each language group bestows a set of color terms to categorize the perceived colors, limiting the number of available color terms in a person's vocabulary. Language allows us to partition the continuous space of color into a manageable set of discrete categories (i.e. basic color terms [16]) [78, 111, 104]. However, there are boundaries between these categories and due to the competing nature of the categories, name assignments become less certain towards the boundaries. Therefore, much of the individual variation in color naming systems occurs at the category boundaries. We transform the data (see 2.3.2) into a form that captures these differences in individuals' category boundaries in order to incorporate as much of the structural nuance into the model as possible.

The WCS *naming task* (see Section 2.3.1) was a forced naming task where each participants was required to provide every color chip with a name regardless of the chip's salience. These regions of low salience are typically on the category boundaries [79, 40]. Consequently, especially when assigning a name to a color chip in these regions, participants rely on their own internal perceptual processing to make a judgment. These visual processes are universal biological features for most people. Since these processes are shared amongst most of the population, we can anonymize participants in the ML application. To do so however, we first abstract away from each particular language by using features of the color space to transform the data into a form that is agnostic of their culture and language.

The stimulus space used in the color naming task is a discrete set of color chips sampled from a Mercator projection of a 3D color solid with its own unique geometry [57]. In this color space, the colors which are closer in proximity will be perceived more similarly. Such a stimulus space is fundamentally different than one where each stimulus is independent from other stimuli (e.g. pictures of dogs) or one that is directly related to each other (e.g. tracking movement of a ball in a video clip). There is both an underlying relationship between color chips in the space yet each chip is a unique stimulus. There is a neighbor-like relationship between adjacent color chips that we use to transform each participant's color naming data into pairwise judgments between these neighbors (see Section 2.3.2). This transformation maintains the participants' original categorizations and abstracts away from their linguistic origin, enabling a language-independent way of studying universal patterns. The BBDP will draw similarities between participants based on their neighborhood judgments.

2.3.1 Data

This paper uses *naming task* data gathered from participants of the WCS [22]. It consists of participants assigning names to each of the 330 Munsell color chips, presented in a fixed, random order. Of the 110 languages surveyed, four languages are omitted from our data set due to data collection and transcription issues. The four languages omitted are: Huastec (Language 45), Mampruli (Language 62), Tarahumara-C (Language 92), and Tarahumara-W (Language 93). Additionally, the 10 achromatic chips (column A0–J0 in Figure 1.5) are omitted from our data set due to their disjointed nature from the 320 chromatic chips. Therefore, "the data" or "our data" will henceforth refer to the collection of *naming task* data for 106 languages (2,552 individuals) on the 320 chromatic chips in the color grid.

2.3.2 Transforming Color Data

To prepare the data for the BBDP, participant color categorizations are converted into an n-dimensional binary features vectors. In their construction, each chip's chosen name is compared to the names of its immediate neighbors. Chips B1–I40 in Figure 1.5 are considered one at a time to be the *reference chip* and its name is compared to the name of its vertical and horizontal neighbors (rows B and I have three while all other rows have four neighbors. The color grid is a Mercator projection of the Munsell color space, so the ends of the rows are considered to be connected or adjacent (e.g. B1 and B40) but the top and bottom rows (B and I, respectively) represent the poles of the 3-dimensional solid. Therefore, chips in rows B and I do not have vertically adjacent neighbors in the north and south direction, respectively.Each element of the binary vector represents one such pair. An index in the vector is given a value of 1 if the two color chips being compared have the same name and 0 if they have different names (see Figure 2.3).

Transforming the *naming task* data yields a set of 2,552 data points each with 2,320 binary

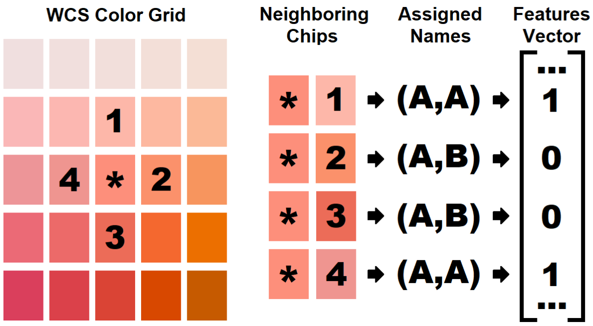


Figure 2.3: An illustrated example of the transformation process for a chip from a color naming system to binary features vector. First, a color chip is selected as the *reference chip* (represented by *). Second, * is compared to chip 1. They are assigned the same name, therefore, 1 is recorded in their features vector's corresponding index. This is repeated for chips 2 to 4.

attributes. Redundant pairs are omitted from the features vector—(chip i, chip j) is included in the vector, but not (j, i).

2.3.3 Model Implementation

We use the Python package BayesPy [84] to implement variational inference methods. It has built-in functions for performing variational Bayesian inference on conjugate exponential family models. BayesPy approximates infinite dimensional distributions, such as the Dirichlet process, by setting the maximum number of clusters K to a value much higher than the number of expected clusters. We initialize $K = 100$ clusters and consistently find the number of resulting clusters $K^* < 100$, indicating that the results are driven by the data and not an upper bound.

The hyperparameters for the BBDP—α (Dirichlet concentration parameter) and β_1, β_2 (beta distribution shape parameters)—are estimated using the random search method. Through one hundred random initializations of the BBDP, optimal parameters are identified by the model with the highest ELBO. Following several random searches, the hyperparameters selected are are $\alpha = 1000$ and $\beta_1, \beta_2 = 0.9$. These are used to generate the results reported in Section 2.4.

Though one of the main benefits of implementing a model using Bayesian nonparametric methods is the ability of the model to freely determine parameters' values throughout the training process, these models still contain variables which need to be exogenously determined. These variables are called *hyperparameters*. Several methods are commonly used in order to search the parameter space for set of hyperparameters which will yield the most "optimal" result, such as grid search, random search, Bayesian optimization, and evolutionary optimization [14, 15]. We chose to use random search to estimate values for our model's hyperparameters.

A naïve search method is employed instead of one which actively searches for an optimum (e.g. Bayesian or evolutionary optimization) because the more simplistic approach was found to be sufficient for the purposes of this study. Hence, random search is used to search the parameter space for an estimate of the optimum. The optimal parameters are the determined by finding the combination which yields the highest ELBO.

Several sets of 100 random initializations were run at a time in an effort to determine general regions of optimality. This revealed a broad pattern. Higher values of β generated higher ELBOs whereas α did not appear to have much of an influence on the value of the ELBO (see Figure 2.4). This finding is consistent with the fact that the model is more sensitive to the beta distribution hyperparameters than the Dirichlet process concentration parameter [97]. Therefore, based on the pattern obtained from running multiple set of initializations and precedence from previous literature, the hyperparameters selected for the model were $\alpha = 1000$ and $\beta = \beta_1, \beta_2 = 0.9$.

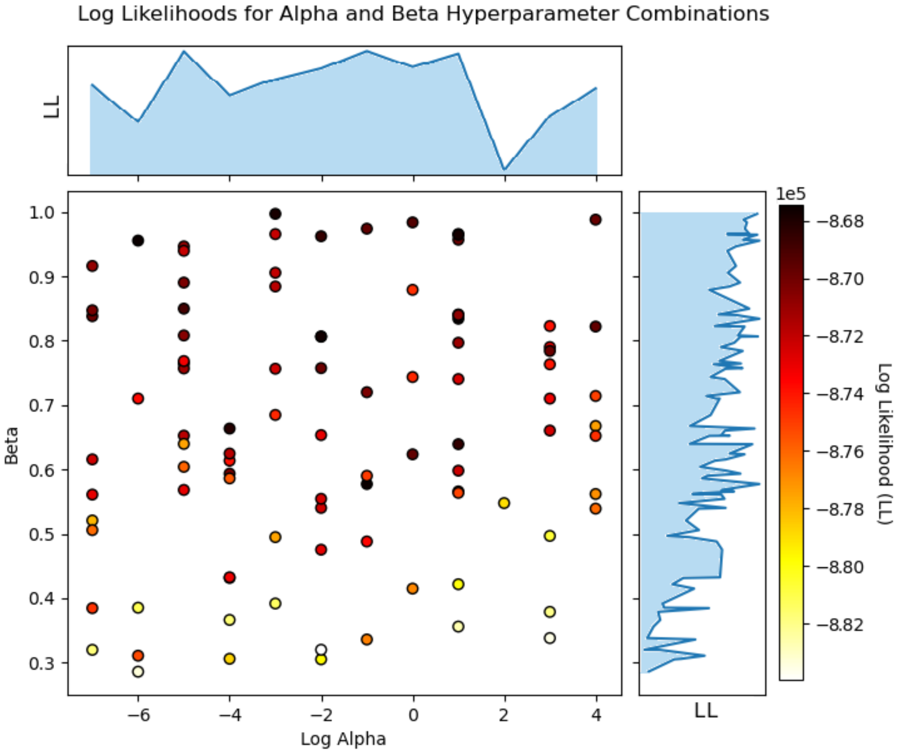


Figure 2.4: Scatter plot representing 100 random initializations of the BBDP. The x-axis represents the α parameter (Dirichlet process concentration parameter) in log units. The y-axis represents the β hyperparameter (β_1, β_2 of the beta distribution). The darker the color of the data point, the higher the ELBO of the algorithm run using those hyperparameters.

2.3.4 Cluster Analysis

Specific and novel methodologies are developed to perform within and between cluster comparisons. These methods are useful for visualizing the resulting clusters and conducting customized statistical and quantitative analysis.

Visualization Using Centroids

The *centroid* map provides a singular representation of each cluster. They are constructed by taking the modal term for each color chip. Doing so gives the best representation of all the members in the cluster (i.e. the distance from each member to the map is minimized).

Given that a cluster can consist of participants from more than one language group it is necessary to first create a translator that maps all color terms into one common "language". This is done by performing k-means clustering over the terms used by all participants in the cluster. Terms in the same cluster are considered equivalent (i.e. refer to similar regions of the color space). This is the method established in [78] with the exception that the gap statistic is exchanged for the average silhouette method [110, 66] to determine the optimal number of clusters. After all participants are translated into the common "language", the centroid is constructed by taking the modal term for each color chip.

Visualization Using Boundary Heatmaps

To reveal the underlying strength of the category partitions, a more informative representation of the resulting clusters is depicted through *boundary heatmaps.* While centroids reveal the color space partition, boundary heatmaps show the strength or level of agreement over the partition. This additional information is imperative because it discriminates between two similar modal maps by revealing the regions within the partition that are most salient to the cluster.

The boundary heatmap assigns a value to each color chip in Figure 1.5 by computing its *boundary probability* (i.e. the likelihood that a color chip is located on the boundary of a color category) introduced in [40]. This represents the strength of the category boundary at that chip.

Schematic Similarity

Schematic similarity (SS) allows within and between cluster analysis by comparing the level of similarity between two participants' naming systems regardless of their language of origin. This analysis is based on the partition itself and, therefore, is not dependent on the names assigned to regions of the color space. Particularly, in the absence of any cultural knowledge of the languages, SS can still assign an objective similarity score. SS captures maximal information by comparing similarity at the participant-level.

SS calculates similarity between two participants by determining the amount of overlap among their corresponding terms. Terms that refer to similar regions of the color space are mapped to each other. Once all the terms each participant are mapped, the metric does the following to compute the amount of overlap:

1. Set the participant with fewer color terms as the *Reference Participant* (RP) and the other as the *Other Participant* (OP). Let n be the number of color terms used by the RP.

2. Find the best term equivalence for RP's Term 1 using *Term Similarity* by

$$TS(T_i^{RP}, T_j^{OP}) = \frac{|T_i^{RP} \cap T_j^{OP}|}{|T_i^{RP} \cup T_j^{OP}|} \tag{2.9}$$

where T_i^{RP} represents the set of color chips that RP named with term i and T_j^{OP} represents the set of color chips that OP named with term j.

Let T^{OP} be the set of terms used by OP. The term used by OP which best matches T_i^{RP} (Term i used by the Reference Participant) is $\underset{j \in T^{OP}}{} TS(T_i^{RP}, T_j^{OP})$.

3. Let M be a set, where each element is a tuple representing the mapped terms between RP and OP. Suppose from Step 2 that $\underset{j \in T^{OP}}{} TS(T_1^{RP}, T_j^{OP}) = k$, then the tuple $(1, k)$ would be added to M.

4. Repeat Steps 2 and 3 for Terms 2, ..., n used by RP (see Figure 2.5a).

5. Once all of RP's terms have a best match, calculate *Schematic Similarity* between RP and OP.

$$SS(RP, OP) = \sum_{(i,j) \in M} \frac{1}{n} \frac{|T_i^{RP} \cap T_j^{OP}|}{|T_i^{RP} \cup T_j^{OP}| + e_{T_i^{RP}}} \tag{2.10}$$

where $M = \{(i, j), (a, b), ...\}$ represents the set of corresponding/mapped terms from RP to OP and $e_{T_i^{RP}}$ represents the error in term i.

Term error is calculated using the following formula,

$$e_{T_i^{RP}} = \sum_{k \in U} |T_i^{RP} \cap T_k^{OP}| \tag{2.11}$$

where U is the set of terms used by OP that did not get mapped to an equivalent term from RP (see Figure 2.5b).

After determining the SS between two participants, the distance between the participants is

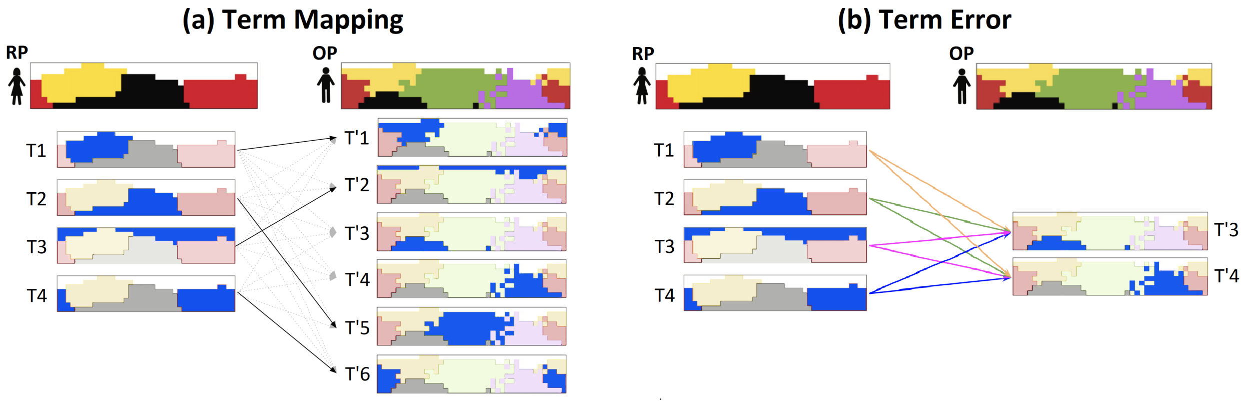


Figure 2.5: Example of (a) term mappings and (b) term error where each term is represented by the blue region in front of T for the RP or T' for the OP. In (a), left participant is RP (n=4) and the right participant is OP. Black arrows represent the term mapping (e.g $T2$ mapped to $T'5$). In (b), two of OP's terms, $T'3$ and $T'4$, are not mapped to any of RP's terms and contribute to the Term Error incurred by each of RP's terms. For instance, the error of $T1 = |T1 \cap T'3| + |T1 \cap T'4|$.

defined as:

$$d(RP, OP) = 1 - SS(RP, OP) \tag{2.12}$$

The distance metric shows the dispersion within a cluster and can also calculate the distance between cluster centroids. (See A for the proof that $d = 1 - SS$ is a metric.)

Group Error

Due to the forced nature of the WCS *naming task*, regions of the color space with low cultural salience likely caused some of the observed variation in the naming systems of participants from the same language. To study within language variation, we develop *group error*. Group error is a novel measure that quantifies the diversity within each language group for post cluster analysis. With this measure, we can investigate the distribution of universal patterns across groups. Group error is a function of the number of clusters a language group is divided into and the distance between those clusters. To calculate the error for that language, participants are considered in a pairwise fashion. The distance between two participants in the same cluster is 0 while the distance between two participants in different clusters is the distance between their respective clusters. Group error is formally defined as:

$$e_g = \frac{\sum_{(i,j)} d_{ij}}{\binom{N_g}{2}} \tag{2.13}$$

where N_g is the number of participants in the group g, $i, j \in \{1, ..., N_g\}$, and

$$d_{ij} = \begin{cases} 0 & \text{if } \ell_i = \ell_j \\ 1 - SS(C_i, C_j) & \text{if } \ell_i \neq \ell_j \end{cases} \tag{2.14}$$

where ℓ_i represents label of the cluster containing participant i, ℓ_j represents label of the cluster containing participant j, and C_i and C_j are the centroids of the i and j's clusters, respectively. Since group error is a novel contribution, we used an established diversity measure—Simpson's Index—to validate it. We found Simpson's Index and group error to have a highly significant correlation of $\rho = 0.9$ ($p\text{-}value \approx 0$). Thus, we have evidence that group error is capturing the intended phenomenon.

Simpson's Index is a measure of the diversity of species within a habitat. In our case, WCS language are considered the habitats and the species are the clusters that the language's participants are assigned to by the model. The measure can be written as:

$$D_s = 1 - \frac{\sum_{i=1}^{R} n_k(n_k - 1)}{N_g(N_g - 1)} \tag{2.15}$$

where N_g is the number of participants in the WCS language group (habitat), R is the total number of cluster labels (species) in the group, and n_k is the number of participants from the group who are in cluster k. Since Simpson's Index is a well-known, established diversity measure, it will be used to validate the novel group error measure, which was developed to take into account specific features of our model.

2.4 Results

2.4.1 Model Efficacy

To determine the effectiveness of the BBDP model for the tasks outlined, it was initially tested with "synthetic" data generated by simulations from [40]. The data consisted of language groups with nearly identical naming systems; therefore, we expected the BBDP to place members of language groups in the same cluster. We found that 84% of languages had a group error of 0 (i.e. all members of the language were assigned to the same cluster), thus demonstrating that the model successfully identified similarities in naming structure and accurately grouped them together.

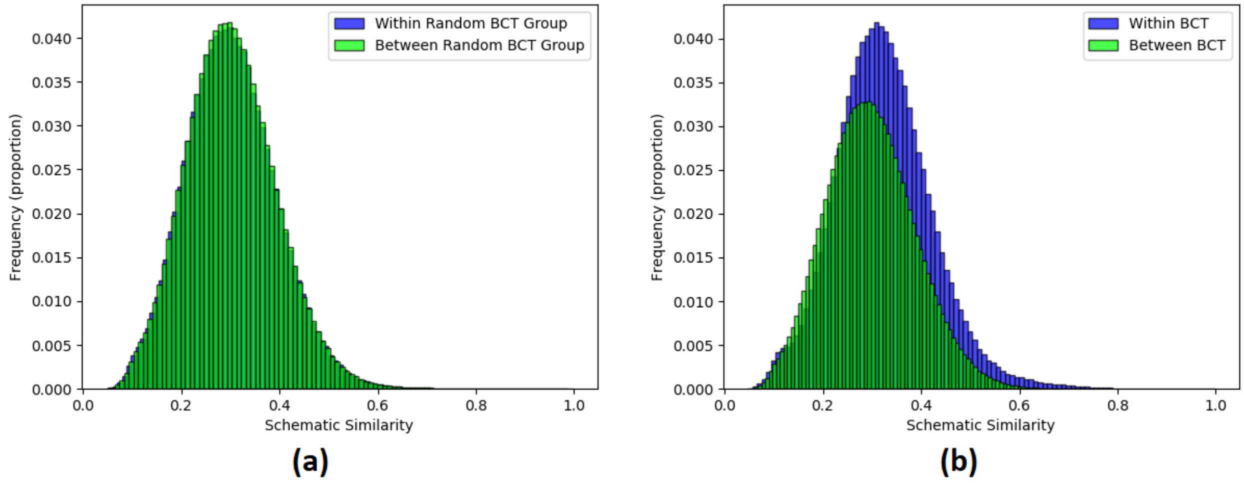


Figure 2.6: Distributions representing within group SS (see Section 2.3.4) and between group SS for (a) randomly assigned psuedo-BCT groups and (b) WCS identified BCT stages.

After BBDP successfully clustered the "synthetic" data, next the task was performed with our data. The resulting clustering were compared with three baseline groups as a point of comparison: (i) a random grouping of participants; (ii) an existing grouping of participants (e.g. by the number of basic color terms (BCTs) identified for the languages in the WCS analysis [69]). We generated SS histograms to compare the level of similarity within and between the aforementioned groups (see Figure 2.6); (iii) a representative subset of participants.

For (i), we randomly assigned participants into 10 groups to represent "psuedo-BCT stages". WCS systems identified to have anywhere from 3 to 12 BCTs; thus, 10 random groups to represent 10 BCT stages. The within group histogram had $\mu_w = 0.30$ and $\sigma_w = 0.092$, and the between group histogram had $\mu_b = 0.297$ and $\sigma_b = 0.092$ (see Figure 2.6a)—the two distributions are almost perfectly concurrent. This pattern suggests that the random groups were not capturing inherent structure among participants.

For (ii), participants were assigned to groups with the same number of language BCTs—e.g. members from languages with 3 BCTs. This grouping is based on properties used in [16, 68] to draw comparison between naming systems. The within BCT group histogram had $\mu_w = 0.324$ and $\sigma_w = 0.098$, and the between BCT group histogram had $\mu_b = 0.297$ and $\sigma_b = 0.0899$ (see Figure 2.6b). This slight positive shift in the within group SS away from the between group SS indicated that the number of basic color terms is capturing some of the similarities among color categorizations.

Comparing the results of the first two baselines groups to the clustering results generated by our model helps evaluate how well the inferred clusters are capturing the structural similarity between participants. Our clusters were found to be better representations of both baseline group with a higher within group SS (see Figure 2.7). The within cluster distribution had a $\mu_w = 0.384$ and $\sigma_w = 0.111$, where as the between cluster distribution had $\mu_b = 0.298$ and $\sigma_b = 0.089$. The more dramatic positive shift in the within cluster distribution away from the

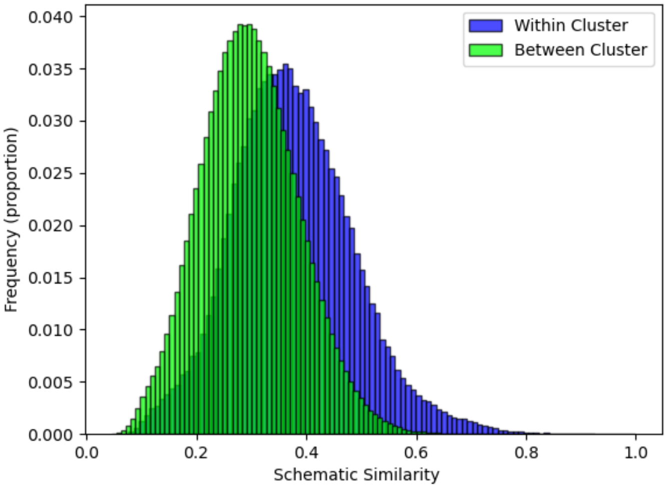


Figure 2.7: Distributions representing within group SS and between group SS for resulting clusters from the BBDP.

between cluster distribution indicates that the model is capturing similarities and nuances in the structure of these color naming systems that was not being captured by either of the baseline groups. Comparing results from clusters to the third baseline group ($\mu_w = 0.404$, $\sigma_w = 0.130$, $\mu_b = 0.302$, $\sigma_b = 0.0919$) showed little difference, further validating the BBDP's performance.

2.4.2 Clustering Results

One run of the model using parameters $K = 100$, $\alpha = 1000$, and $\beta_1, \beta_2 = 0.9$ yielded a total number of 88 inferred clusters. Of these 88 clusters, 85% of participants were placed in 18 clusters, visualized by centroids and boundary heatmaps in Figure 2.8. The remaining clusters each contained less than 1% of the data set and were removed from subsequent cluster analysis.

The average SS of all cluster means was 0.342 and the standard deviation was 0.067. In comparison, the average mean SS for clusters and language groups were comparable, but the average language standard deviation ($sigma = 0.080$) was higher than for the clusters.

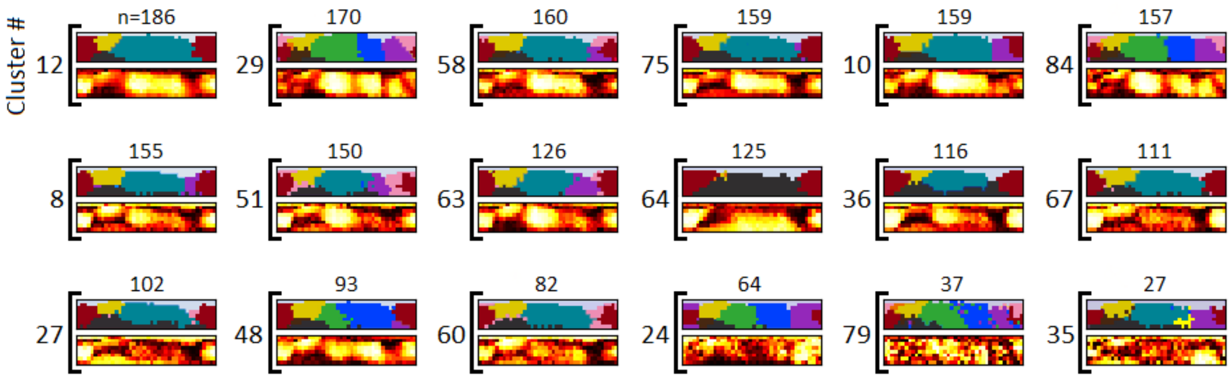


Figure 2.8: Cluster representations for the 18 analyzed resulting clusters with number of members (n=) shown on the top and cluster number (#) on the left-hand side of each group for reference in Figure 2.9. Clusters are visualized through their centroid (top) and boundary heatmap (bottom). Colors in the centroid are used only to distinguish lexical terms and loosely denote the hue of the underlying region—they are false colors. In the boundary heatmaps, the darker the color, the higher the likelihood that the corresponding chip is located on a category boundary.

The smaller standard deviation of resulting clusters inferred that the resulting clusters from the BBDP were more tightly clustered than the language groups. This provides further evidence that the clusters are better representations and are better capturing the similarity in structure of the naming systems.

Examining each cluster, we found similarities across languages (i.e. multiple languages were represented) as well as diversity within languages (i.e. participants from the same language did not necessarily get assigned to the same cluster). The average group error across all languages was $\mu_{e_g} = 0.314$ and was found to be statistically significant ($t = 35.40$, *p-value* ≈ 0) meaning languages typically could not be characterized by a single naming pattern. This is corroborated by the fact that participants from a language were found in 6.83 different clusters on average (median=7, mode=7, 8). The range of this value, however, varies from 1 to 13 (see Figure 2.9), highlighting that languages have varying levels of diversity. Group error and the number of clusters that the languages get divided into are positively correlated ($\rho = 0.712$, *p-value* ≈ 0; see Figure 2.9). Observing the distribution of the languages across the resulting clusters provides insight into the structures (and possible

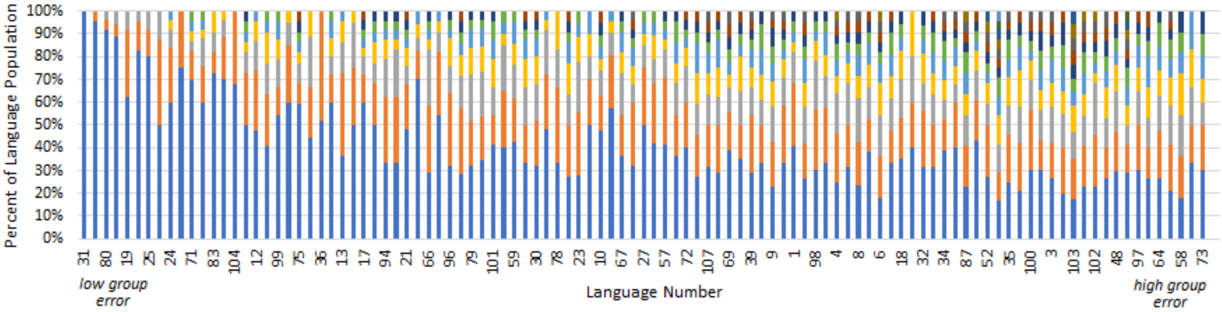


Figure 2.9: Distribution of each language across the inferred clusters ordered along the x-axis from lowest to highest group error. Each bar is a language group, each colored segment represents the clusters the members were placed in, and the size of the segment is proportional to the number of participants in the corresponding cluster. For example, all members of Language 31 (far left) belong to the same cluster while Language 73 (far right) has members dispersed among seven clusters.

evolutionary processes) present in the language at the time of collection. For example, some languages, may be characterized by a single, prominent color naming system (e.g. Language 74 [Múra Pirahã] has all but one member belonging to cluster 36, see Figure 2.10a) or by a few strong systems (e.g. Language 85 [Seri] has members dispersed between three main clusters [12, 10, and 29] with two in other clusters). The low group error of Language 74 points to high levels of similarity and cohesion, implying that conventions associated with its lexicon are well-established within the population. This could be an indicator that Language 74 has reached a point of stability within its evolutionary stage. Conversely, the diversity present in Language 85 points to poorer lexical consensus among its speakers. This diversity could be influenced by different factors: (i) gender differences [30], (ii) weak notions of color [74, 79], or (iii) a transition from one evolutionary stage to another [78] (Cluster 12 contains the individuals without a "pink" term whereas participants in Clusters 10 and 29 possess a term for "pink"—this could point to the introduction of a new term into the system). While the methods described here are certainly useful in examining broad trends and identifying points for further analysis, the linguistic and anthropologic tools to draw such conclusions are beyond the scope of this paper.

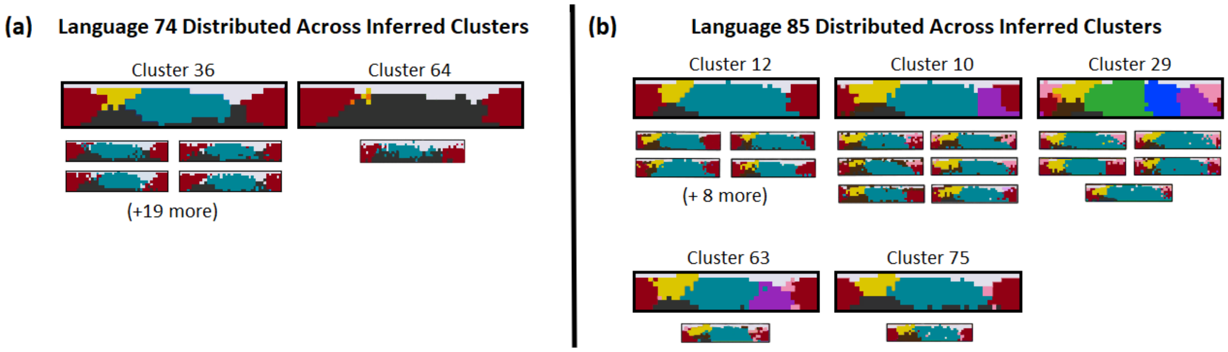


Figure 2.10: Distribution of language participants across clusters from a language with (a) low group error and (b) high group error. Larger colored rectangles represent the centroids of the cluster. The smaller colored rectangles below the large ones are the name assignments for each language participant found in that cluster. As in Figure 2.8, the colors depicting the categories are false colors.

2.5 Discussion

The primary aim of this paper is to provide an initial demonstration of the application of Bayesian nonparametric models to color naming data. Building on previous work [77, 78] which used parametric clustering algorithms, we investigate underlying features of color categorization systems with a particular nonparametric method, BBDP. This novel approach has findings which are in agreement with the existing literature.

The largest commonality between our results and previous conclusions is that the 110 WCS languages can be classified into a small number of patterns. This fact coupled with the widespread nature of these clusters across unrelated languages suggests the existence of universal processes which govern color naming systems. Our inferred clusters share resemble motifs previously identified by [78]—namely, they are representations of shared patterns across individual naming systems. Another finding consistent with past research is the significant amount of diversity that exists even among individuals from the same language [78, 133, 30, 40]. Group error finds that most languages have multiple patterns present (∼7), highlighting that diversity cannot be predicted by language membership alone.

This paper extends the existing literature by further abstracting away from both linguistic and algorithmic assumptions. As mentioned previously, one of the main draws of Bayesian nonparametric models is the ability to let the data determine the complexity of the model. The most prominent work to date which uses parametric ML algorithms on the WCS [78] identified 3–6 motifs but pointed to the need for a different approach to capture "minority motifs" that are rare but resemble evolutionary stages proposed by [16, 67, 69]. Using BBDP yields 18 prominent clusters with some having near equivalence to the motifs, with the remaining clusters possibly being these "minority motifs". Knowing the distance between the resulting clusters in conjunction with the boundary heatmaps presents a more nuanced view of the data.

A notable contribution made by this paper is the inclusion of distance when analyzing the spread of the resulting clusters. While previous literature has looked at the distribution of "motifs" across the WCS language groups, there has not been any analysis to see how far these resulting clusters are from each other. By analyzing not only the distribution of clusters but also their proximity to one another, we can uncover more information about the structure of each WCS language's population naming strategy.

The successful application of the BBDP using the transformed binary data (see Section 2.3.2) shows that the "color neighborhood" is a meaningful criterion by which to evaluate structural similarities between individuals. Results demonstrate that using local boundary information can replicate and build on past findings. Roberson et al. asserted that the determining feature of color categories were boundaries which varied across languages [**?**], not universal focal colors as others had claimed [70, 69, 47, 48, 106, 107]. We similarly find that category boundaries are a deterministic feature of these systems, but unlike Roberson et al. we suggest that these boundaries are influenced by universal factors perhaps in addition to more language-specific phenomena. Given the recent developments in tools for detecting color category boundaries [28, 40], future research can further explore the properties of these boundaries.

2.6 Application

The results outlined in this paper can lead to new investigations for color researchers while methods provide a use case for the application of ML techniques for social science topics.

A significant portion of the study of color naming conventions and its evolution has been informed by the World Color Survey data [69, 78, 11, 136, 40]. The WCS gathered data from a diverse set of monolingual, preindustrial language groups, making it particularly useful for the study of color naming. However, the purpose of the WCS was to test hypotheses originally posited in [16] and was therefore collected with theoretical assumptions in mind (see section 2.3.1). This execution grounded the data in Berlin and Kay's theory, limiting the number of alternative theories that can be tested. With this limitation in mind, our paper customizes advanced ML methodologies to circumvent the theoretical constraints and shows how to investigate universal patterns without reliance on linguistic theories or cultural knowledge. The resulting clusters revealed 18 universal patterns and their salient features. Using our findings, color researchers can investigate common features (e.g. geography, age, gender) among the clusters to determine what factors underlie the color naming schema of the cluster (e.g. younger participants get clustered together). Additionally, features of these clusters can help build new hypotheses that will lead to future empirical studies and guide new data gathering efforts.

The methodology outlined, on the other hand, can be used for many other social science topics. For example, ML techniques similar to those used in this paper can identify the number of natural religious, political, or cultural groups and their key features based on participant responses—i.e. Muslim and Jewish participants may belong to the same cluster. Similar to methods used on the WCS, starting with a set of religious groups (e.g. Methodists, Mahiyana Buddhist), BBDP can identify latent groups with similar beliefs. Insights from such studies can be useful in identifying extremist beliefs, understanding attitudes towards marginalized groups, or predicting when groups will split. Whereas the study of religious, political, or cultural beliefs gets artificially divided into topics and studied by various departments, each with different aims, ML allows researchers to study them holistically. It does so by allowing the structure and patterns underlying the data to guide investigation rather than theory as the main driving force. Here, we provide a roadmap to accomplishing this task.

2.7 Conclusion

This paper successfully demonstrates the application of advance machine learning techniques (e.g. generative Bayesian nonparametrics) to study social phenomena based on universal features. While these models have been widely applied to text and image processing where number of natural categories is unknown but of great interest, its application in the social sciences has been sparse. Using the second approach to universality, we base the ML model for color naming on universal features of linguistics, biology, and the color space to uncover extant patterns in WCS data without relying on cultural knowledge of the language groups. Specific customization for this goal include transforming the data into neighborhood judgments based on the topology of the stimulus space and developing appropriate mathematical techniques to study results. We show that BBDP is an appropriate model to uncover universal naming patterns and find our results to be in agreement with past work demonstrating that there exist universal tendencies among color naming systems worldwide.

Social scientists interested in studying universality in social problems can deploy similar methods used in this paper. For other applications, there are a wide range of ML algorithms available. Additionally, those interested in further investigating the WCS can use the 18 universal patterns in this paper as a starting point. Though this paper provides initial insights into the structure of naming patterns found across the WCS languages, it does not evaluate *why* these structures might be present. Therefore, researchers with the appropriate cultural knowledge and tools can delve deeper in understanding the societal and cultural processes present in each language. This further exploration could answer questions such as: Does a person's role in society (e.g. hunter versus gatherer) affect their color categorization? Do men and women have different vocabularies? Do languages in proximal geographic regions share structural similarities? Does this apply also to languages with the same temperate environment?

Another avenue for future research is to compare our results against other Bayesian nonparametric models. The model described here serves as a preliminary approach which, given its success, provides an avenue to test the data using more complex algorithms. Testing these complex models could be useful in uncovering other facets of color categorizations not revealed by this particular approach. Additionally, using such algorithms would result in fewer small, outlying clusters which would diminish the number of data points removed from analysis. Though there are points of expansion, the results presented here are a satisfactory introduction of nonparametric models to color naming.

Chapter 3

“How ’me’ becomes ’we’: A Computational and quantitative approach to ideology

Shared beliefs or ideologies have been long theorized to shape group attitudes and behavior. This paper utilizes Cultural consensus Theory (CCT) and machine learning (ML) techniques to help define and identify systems of beliefs or ideologies present in groups. Here, ideology is thought of as a social convention that creates within-group cohesion by establishing norms that dictate “right” and “wrong” actions and beliefs, ideas that create collective meaning (e.g. holy, fair), and tacit rules for interaction with inside group members and outsiders. Such a conception while broad, provides a starting point for studying ideology organically (i.e. theory free). In this study, we ask a participants with diverse memberships to answers true or false to a range of cultural questions on behalf of their groups. We find that of the four major clusters, two are moderates with little political consensus, and the other two are traditional conservatives and liberals who oppose conservative beliefs.

3.1 Introduction

Ideology is a difficult topic to research because its definition varies between thinkers, researchers, and fields. One common way to think about ideology however is as a shared system of values, beliefs, and attitudes that helps a group make sense of reality and prescribes how it should be or work. The key part of this conception of ideology is about it's "shared" aspect. Ideologies do not belong to individuals but rather to social groups, particularly large ones (e.g. political, religious groups). This paper studies ideology's shared system as a social convention. Here, a social convention is thought to create within-group cohesion by establishing norms that dictate "right" and "wrong" actions and beliefs, ideas that create collective meaning (e.g. holy, fair), and tacit rules for interaction with inside group members and outsiders. By applying well established quantitative methods not utilized for the study of ideology, I measure group cohesion, common knowledge, and how they differ from other groups. Social conventions such as shared beliefs or ideologies influence group attitudes and behavior. Understanding how ideologies form, evolve, and influence groups can provide powerful insights incentivize cooperation between groups, using insights from existing work through targeted social interventions. This can be particularly useful in identifying extremist beliefs, incentivizing prosocial behavior, changing attitudes towards marginalized groups, increasing between-group cooperation, to name a few.

This paper examines the structure of ideological groups in a novel way. While prior work has focused on political and religious beliefs separately, there has been relatively little work on overarching ideologies. Using Cultural Consensus analysis allows us to discover a groups' cultural "answer key" or ideology. This method has been extensively in a wide variety of domains: in studying medical knowledge and beliefs in anthropology [135], in extracting information from eye-witness testimonies [131], in inferring judgment of personality traits in social networks [1], in evaluating interpersonal agreements on psychological concepts such as behavior [99], and cultural concepts such as What does it mean to feel loved [49].

In applying it to the study of ideology, we look at groups that are commonly thought of as having ideologies and apply cultural consensus methodology to these groups so that we can measure group cohesion, common knowledge, and how they differ from other groups. We can later move on to other types of groups (e.g. k-pop fans) and see whether they are structurally similar in common knowledge to these other groups. What is novel in my approach is that not only do I anonymize the participants in the analysis we also anonymize the groups in comparative analysis, thus we do not know in our mathematical analysis using ML which anaomymous person comes from which group and we rediscover the groups by patterns of their questionnaire behavior. What this allows for is to understand in an objective way the universal properties of ideological groups versus other types of groups. What thing that this novel approach will give us is a way of finding latent ideological groups that are recognized as such in the culture but that has all the properties of an ideological group.

In this bottom up approach, we do not start with the idea of ideology but rather with looking at ideological groups and then finding which beliefs are giving rise to the groups. There may be several ideas about ideology but at the group level. Doing so allows us to bring scientific social science methods to ideological groups based on common features.

3.1.1 Cultural Consensus Theory

To identify ideology groups, this paper uses cultural consensus theory (CCT) [12]. CCT is best suited for social phenomena where there are inhomogeneous respondents as well as a latent cultural "answer key" that is unknown to the research a priori. That is, when the topic of study is unknown to the researcher making it difficult to ask relevant, objective questions that would lead to gaining objective knowledge about a group. Anthropologists, linguists, and social psychologists that attempt to investigate objective knowledge in cultures different from their own face such challenges routinely.

The CCT framework assumes informants come from a coherent cultural unit and each possess knowledge about their common culture. Of course the informants may vary in their competence, so some of their responses may differ with their differing levels of expertise. The problem the researcher faces is to aggregate the data to reach normative conclusions about the culture. By asking participants to answer each item based on "Your group would agree that...," CCT analysis is helpful to reveal the different levels of question relevancy: culturally salient items are expected to have a high degree of consensus while the opposite is true for low salient ones. CCT as an information pooling technique looks to patterns of agreement or consensus across items to make inferences about each informant's level of ability or competence. Those with a higher differential competence are identified as *experts* and their responses are given more weight than non-experts. Mathematically, experts are individuals whose responses most corresponds to others in the group (i.e. the person closest to the “cultural answers” of the group). To summarize, CCT allows for 1. determining whether observed variability in level of knowledge is cultural; 2. measuring each person's cultural competence; and, 3. revealing the culturally correct answer or knowledge.

Other models it derives from:

1. **Signal detection theory:** general detection model where it estimates respondent "ability" and guessing bias parameters (types I and II errors)
2. **Latent Structure analysis:** Produces the “answer key” of a cultural group by interchanging respondents and items in the analysis. CCT is structurally isomorphic to the two-class latent structure model.
3. **Item Response theory:** Uses same data structure and rather than measure the latent properties of the respondents, CCT measures the latent property of items. When a researcher lacks “true” knowledge of categories, using CCT, the level of salience or cultural relevance of an item is estimated to aid conducting objective research – minimizing researcher bias.

4. **Condorcet jury trial problem:** General Condorcet Model (GCM) experts have more weight
5. **MAP and MLE Bayesian estimation:** Used for identifying number of cultures present in the data.

CCT is unique in that it deals with an unknown cultural answer. Past models either dealt with known answer keys or discovered latent groups with no answer key. psychometric test theory models for example, estimate a person's latent ability parameter based on their test performance and item difficulty but in such cases, the researcher knows the unitary answer key. There are numerous social situations where researchers lack such objective knowledge. On the other hand, in sociology and political science, latent structure analysis (LSA) models are used to discover opinion structures of respondents based on the number of homogeneous latent groups present in the data. While a powerful tool for gagging public opinions, for example, the model does not aim to discover the underlying cultural answer that we expect to see in many social groups– we aim to understand the ideology of an ideological group.

There are other models that evaluate whether agreement in a population is significant (e.g. Binomial test, Friedman test, or Kendall's coefficient of concordance). The advantage of CCT over such models is that is also estimates the "true" answers as well as the level competence of each person. Because CCT derives this answers by "weighting" individual responses based on competence prior to aggregation, its aggregation is more accurate that the combined responses of individuals. This is a key advantage of CCT as a data aggregation tool.

3.2 Methods

Question development, testing, and data gathering constitutes a majority of the methodology for this study. The approach to clustering in this study mirrors the methodology described in Joe et. al. [62] (see Chapter 2).

3.2.1 Data

The data for this study is gathered online from Amazon Mechanical Turk (MTURK) and undergraduates at the University of California Irvine (UCI). From the total number of respondents, $N = 114$ passed the exclusion criteria and are included in this analysis. The data consists of *True/False* responses to various types of culturally relevant questions (see below) for a total of $Q = 64$ questions. The data also consists of background information such as demographic data.

3.2.2 Question development and testing

For this study, we assume that groups who identify with or adhere to a specific doctrine or worldview–religious, political, cultural–have an ideology, and our task is to understand what types of ideas, rules, or social structures, create and maintain these groups' ideological content. We identified the following groups as possibly possessing unique ideologies:

To this end, we have developed five types of questions, each with a different logic (in list below), and ask participants from the groups above to answer *on behalf of their group* (not in this order):

1. General, US specific, cultural beliefs

Religious	Political	other
Christian (multiple groups)	Independent	Atheist
Muslim	Republican	Agnostic
Hindu	Democrat	Feminist
Buddhist		Vegan
Jew		Vegetarian
		LGBTQIA
		Yoga
		Environmentalist
		k-pop
		Racial

Table 3.1: List of "ideological" groups surveyed

2. Beliefs about concepts from other groups
3. Ideas about in-group superiority or closeness to "truth"
4. Ideas about out-group inferiority or distance from "truth"
5. Personal benefits from being part of the group
6. Control questions taken from a past study on feeling of being loved ([49])

Over the course of two years, questions were developed and tested on MTURK and UCI participants (total $N = 1484$). Using CCT pack, we were able to assess the difficulty (i.e. cultural relevance) of each item. This aided in selecting the final set of questions such that there was, to the best of our ability, an equal amount of difficulty between groups on each question type.

3.2.3 Model implementation

The CCT model implemented in this study is the implementation of the R package *CCTpack* used for advanced model-based analyses of questionnaire data. This package is based on mathematical publications in the CCT domain –i.e. the General Condorcet Model (GCM)[5], the Latent Truth Rater Model (LTRM)[6], and the Continuous Response Model (CRM)[7]. Respectively, these models are applicable to dichotomous/binary $(0, 1)$, ordinal $(1, 2, ...)$, and continuous data. This package has the ability to detect the consensus answers of the respondents, latent subgroups in the data, their differing consensuses, the expertise of each respondent, their response biases, and the difficulty of each question by parsing the variance of the data.

For the Bayesian nonparametric clustering, we use the Python package BayesPy [84] to implement variational inference methods. By assigning a value much higher than the number of expected clusters to K, BayesPy approximates the Dirichlet process [62]. In this study, we initialize $K = 30$ clusters and consistently find the number of resulting clusters $K^* < 30$, indicating that the results are driven by the data and not an upper bound.

3.3 Results

Based on CCT, participant responses were consistent with the findings in [49] for the control questions. Only one group of sport enthusiasts had a significantly higher level on consensus surrounding "feeling loved" when attending sporting events or playing sports (see section 3.3.1).

3.3.1 Clusters

The respondents were asked to complete a section of the survey featuring binary responses about their group's position on various topics. Each respondent was asked to answer these "for a religious, political, or cultural group [they identified]" or for their peer group, if no strong identification was felt. The binary data was clustered using Bayesian inference on a Bernoulli mixture model. Group membership probabilities have an uninformative Dirichlet prior. Inference on the survey data produced 6 major clusters or 92% of the analyzed data points. If individuals are said to "belong" to the cluster which has a highest assigned probability. The four largest clusters are the most informative. Cluster **17** and **14** have $N = 28$ members each, cluster **4** has $N = 23$ members, and cluster **8** has $N = 15$ members. The names the *eclectics*, the *sunburns*, the *traditionalists*, and the *pragmatics* have been given to each cluster respectively.

Between Cluster differences

In order to analyse the differences between the clusters, an absolute difference (BD) measure is deployed. We have an NxQ matrix of responses where each column represents one question which is a $1xN$ binary vector with 0 representing *false* and 1 representing *true*. We define BD as follows:

$$BD = |\mu_{outgroup} - \mu_{ingroup}|$$

where $\mu_{outgroup}$ and $\mu_{ingroup}$ are the average value of the binary vector (i.e. proportion of true answers).

Though the small size of the data set and the high number of possible membership options limits the amount that can be inferred, using BD we see some notable trends that point to the method's success:

The two kind of moderates

Both Cluster 17 and Cluster 14 are moderates. Both clusters are largely Atheists, Agnostics, and Christians (with a few Buddhists) who almost all identify themselves as "slightly liberal" or "Very Liberal" and are either "Independent" or "Democrat". The interesting difference is beliefs in supernatural forces. Here's what differentiates the eclectics from the sunburns:

One possible explanation is that the eclectics are %71.4 female with many more participants

Only the eclectics	Mutual beliefs
Divine force governing lives	Any good person can go to heaven
Animals have souls like humans	Government cannot be trusted with too much power
Reincarnation	Socialism is a more fair economic system
Astrology for fortune telling	Fossils can be millions of years old
Pets can go to heaven	Group membership brings purpose
Karma	Other liberal ideas (e.g. abortion, LGBT rights, etc.)
Feel loves playing sports	
Feel loved attending sporting events	
Feel loved when solving difficult problems	
Feel loved when the sun is shining	
Group membership to prevent loneliness	

Table 3.2: Group differences between Cluster 17 and Cluster 14

between age 18 - 24 while the sunburns are equally male and female, in their 30s to 50s in age. In fact, the average age of the sunburns is almost two decades older.

Aside from being moderates, the strongest difference based on the BD of items for Cluster 17 is that they agree on not favoring sunshine, sporting events, and solving difficult problems. While similar, the BD analysis for eclectics reveals that their unique belief in spirit for animals, reincarnation, astrology, and love for sporting events is what sets them apart. Neither of these groups have strong religious or political beliefs.

The conservatives

Cluster 4 captures some dimension of conservatism. They include all of the Republicans and almost all Religious group respondents belonging to various Christian denominations (with two Agnostics and one Muslim). The Democrats in this cluster identify as “Slightly liberal”. This groups (“the traditionalists”) is what a general idea of a conservative group would be. Their distinguishing beliefs are that:

- There is a divine force governing our lives
- The American dream
- Pro capitalism

- Religion answering important life questions
- Believe Feminist ideas negatively impact society
- Positive mental health for LGBT living in religious areas
- Freedom of speech under attack
- Reincarnation
- Their group membership brings them closer to the truth, provides a meaning and direction, has social perks (i.e. jobs, friends, spouses), and prevents loneliness

From all clusters, *the traditionalists* are the only ones that have a strong sense of in-group benefits and group meaning.

The liberals

Cluster 8 ("the pragmatics") is largely Democrats and includes several Cultural group respondents (environmental, feminist). Almost all members self identify as "Very liberal" with a few "slightly liberal" and one "Slightly conservative", with almost equal number of male and female members, in their 20s and 30s. This group is the anti-spiritual bunch with equally no beliefs in traditional Christian nor new-age supernatural ideas. What distinguishes them most based on BD analysis is the following:

- There is no divine force in the world
- Animals don't have souls
- No reincarnation
- No heaven
- Generally anti religious ideas (i.e. religion as a positive force)

- Their group membership is not for companionship –or to ward off loneliness
- Their group membership does not provide them with meaning or direction in life
- Their group membership does not lead to living a righteous life

What is common to this cluster is less of a set of political beliefs but rather an antagonistic sentiments towards out group (particularly conservative) beliefs.

Between Cluster similarities

Across all groups, they reported their groups believing in the following ideas at an equal level –i.e. no between group differences:

- Governments cannot be trusted with too much power
- Eating meat is ethical
- Abortion is a complex problem with no clear solutions
- Intelligent parents give birth to intelligent children

3.3.2 Clustering robustness

Given the "hard" clusters obtained for alpha in 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2} we wish to determine the consistency of Bernoulli mixture model process. To do that a metric called the variation of information [93] was employed. Variation of information can be thought of as a metric from information theory related to mutual information, but with the property that it has the triangle inequality (so it is an actual metric).

For two partitions X, Y on a set A, we denote the variation of information by $VI(X;Y)$.

It has the properties:

alpha	1e-5	1e-4	1e-3	1e-2
1e-5	0	3.1516	2.9191	3.0997
1e-4	.	0	1.8615	1.6556
1e-3	.	.	0	1.5867

Table 3.3: Level of variation of information between levels of alpha

$VI(X, Y) <= \log(|A|)$

and

$VI(X, Y) <= 2\log(K)$

where K is a maximum number of clusters.

In this case, with 113 data points in the survey, the maximum VI is $log(113) = 4.73$

Observe the table of VI between alpha-alpha pairs:

Notably the initial choice of alpha (10^{-5}) gives a relatively high distance from the clusterings obtained by other values of this hyperparameter. This seems to result in part from the largest cluster with that alpha being made up of a population that is split between two clusters in the other cases. There is still some shared property among the respondents so clustered as they are largely split between two groups that are consistently clustered with the other choices. For every other pair of alphas the largest clusters have very high overlap.

There is no consistent trend towards evenness of cluster sizes as alpha is varied, though the largest alpha tested (10^{-2}) is visually the most even. In every case the largest cluster is between 1.5 and 2.3 times as large as the second biggest, and the 6 largest clusters contain more than 92% of all data points. There is a general trend for certain groups to be clustered together (e.g., most Christian respondents). More analysis must be performed to identify the patterns of answers that drives the more consistent groupings. See appendix B for graphs showing the distribution of cluster sizes.

3.3.3 CCT and Bayesian testing

The CCT implementation in this paper uses the Bayesian Hierarchical clustering mechanism which differs from the Dirichlet processes in several ways. To verify the robustness of CCT and the resulting clusters, both methods were used. For more than 86% of the time, participants were assigned to the same clusters.

3.4 Discussion

A powerful aspect of CCT is in its ability to help researchers ask objective questions about the group. Here, rather than beginning with a preconception of what an ideology is, we allowed the ML techniques to identify four ideological groups with unique sets of ideas and beliefs. Interestingly, the Traditionalists beliefs were in the affirmative, that is to say their central ideas were about the existence of things. This contrasts with the Pragmatics whose core ideas was in opposition to the Traditionalists. The largest groups–and most participants–however, were moderates. The Sunburns had little political, religious, or cultural beliefs but similar to the Pragmatics, had beliefs in opposition to some commonly held cultural ideas. The Eclectics on the other hand had strong consensus around general spiritual and supernatural ideas.

The high amount of agreement across political, religious, and cultural ideas points to the importance of correctly identifying salient, unique dimensions that set these groups apart.

With the success of the methodologies outlined in this chapter, a large scale study can be conducted. Lack of data was the major shortfall of this study. Given the number of groups that we were hoping to include, there was little to non cultural groups that answers the questions. Future studies will target online groups directly and recruit members.

Chapter 4

Who wins: Applying chess ranking system to the ordering of sins

We presents a novel exploration of religious ideology. Using rank order data of moral concepts (i.e. sins) allows for a deeper exploration of religious beliefs. Over the course of two studies, we utilize Elo rating system to aggregate rank ordering of sins and employ k-medoids clustering to uncover universal patterns across groups. The first study finds that Elo gives a stable group rank order of sins with incomplete data (i.e. participants order subsets of sins rather than the entire set). The second study finds that Christians answering on behalf of devout Christians were able to do so where Atheists failed. The data for these studies were gathered using online surveys. Sin is a particularly useful concept: it is universal across religious groups, it is universal outside of religious groups as a basis for secular morality and law, and it is a classification that is universally known. While the word sin resonates most with religious doctrine, taking the broader concept of sin as any moral transgression exists in any group, be it religious, political, or cultural. In fact, many sins as outlined in religious texts are encoded into law and have great cultural significance. For example Judeo-Christian values underpins America's politics, law and morals.

4.1 Introduction

Ideology can be thought of a system of beliefs, rituals, and norms. While there are vast disagreements as to what constitutes an ideology, religion is typically thought of as an ideological archetype. Religious belief and behavior is a uniquely human experience with evidence of it found in all cultures from past to present [105]. Archaeologists have found evidence of religious ritual among hunter gathers and even traced religion back to our earliest Sapiens progenitors ([126] [94]). Despite secular pundits predicting the demise of religion for centuries, it endures and flourishes worldwide, especially outside of the Western world [119].

Diversity of ritual, dress, diet, and beliefs makes religion into an interesting case study of ideology. Importantly the substantial overlap and universality of its moral code allows for easy comparison. As a system of beliefs, religion is able to generate existential meaning, provide purpose, and sanctions certain behaviors amongst its adherents. Common to all beliefs, sexual chastity or purity is virtuous, prayer is important and powerful, and spiritual discipline such as fasting or meditation are prized activities. But beyond supernatural reward, members who uphold and support these beliefs are able to secure cooperation with other members. Central to devoutness (i.e. adhering to such moral code and beliefs) is the concept of moral transgression, or sin. While there are some differences amongst religious groups around the concept of sin (e.g. *original sin* in Christianity vs. *karma* in Buddhism), the majority of actions that are classified "sinful" are in fact universal across groups. Typically actions that harm co-religionists such as theft, lying, adultery, unlawful killing, greed, pride, envy, etc. are categorized as sinful. Sin as a universal moral concept provides us with an easy case study to analyze group differences and similarities. So while Christians and Muslims would agree that deceit is immoral, they may disagree whether it is more immoral than wrath. The assumption of this paper is that while members of religious groups would agree with which actions are sinful, when asked to order sins from worst to least worst, they may disagree as to which sin is the gravest.

With a considerable overlap of virtuous deed and misdeeds between groups, how do we differentiate between groups? In the first study, a methodology for using rank ordering of moral concepts to understand between and within group differences in presented. this study shows the success of the Elo system as a methodology. Here, we have two groups of participants, each ordering a different set of items. There are three common items in each list which enables for the Elo system to compute the total ordering of the sets.

The second study serves as an exploration of how rank order data can be used to examine how well-known certain beliefs are. Atheists and Christians are asked to order the seven deadly sins as if they were devout Christians and their results were compared to Christians answering on their own behalf.

There are three properties of sins that make a great concept to stud belief systems:

1) It is a universal concept across religions groups

2) It is a universal concept in secular society (i.e. cultural, political, etc.)

3) it is a categorization system and a great majority of sins are shared across groups

Given its universality across religious and secular life (first two properties), we are able to ask individuals with a wide variety of group affiliations about sins. But the kind of questions that would result in a meaningful exploration is influenced by the third property mentioned above. The large over lap in sinful actions across religious and secular groups, simply asking what is sinful or not may not provide insight into differences in beliefs across these groups. Instead, we ask participants to rank order sins from worst to least in an effort to uncover underlying structures of beliefs inherent to groups.

4.2 Moral transgressions or sin

Cultures and societies throughout history have adhered to some form of divine law and viewed transgressions against this law as a grave act or sin. Typically, something is a sin when it is manifested in an action or word although, certain thoughts can be deemed sinful as they can lead to sin. Broadly, there are two types of sin: personal and collective. Sins that include selfish, shameful, harmful, or alienating behaviors perpetrated by an individual can be thought of as personal sin. While sins that are inherent to humanity are collective sin. An example is *original sin* which is central to the Christian faith. Here, "the fall of man," Adam and Eve, from the garden of Eden is due to their disobedience of God's command, leading to the original sin that is then inherited by all decedents of Adam and Eve –i.e. all human beings. This type of collective sin is seen across religions: Pandora opening the box Zeus instructed not to, African bushman making fire that separates humans from animals, and Australian Aborigines defining the sun mother. Such sin is an integral part of many creation myths and serves as an explanation for the state of the world as well as the pain and suffering humans experience. Many creation myths do not rely on collective sin and for the scope of this paper, personal sin (in Christian terms, "Actual" sin) is the only sin of interest.

Importantly, and immediately, sin is tied to beliefs around reward and punishment. A universal idea in religious beliefs is that one's level of adherence to divine law and abstinence from sin determines the outcome for one's life (or afterlife): Supernatural Deity, deities, or cosmic forces reward devoutness commensurate to one's moral accomplishment in this life. But another aspect of punishment takes place in this life and by the hands of religious authorities. Penalties of sins can take the forms of ostracizing, banishment, corporal punishment, punishment of family members, imprisonment, and even death.

From another perspective, sinful actions or thoughts can be thought of as a barrier to ones' connection to the divine or to one's best self. In this light, sins are "bad", not because of punishments or rewards one receives but rather as roadblocks to all that is "good" or true.

Comparatively, there is much variety in the concept of sin. In Abrahamic religions deviation from their set of commandments or rules is sin while a person's ignorance or inability to achieve true self-expression is sin as seen in ancient Greek, Asian, and African societies. The gods also differ in their level of punishment for sin in these religions. The more active, personal gods stand as judge while the detached creators typically leave human to tend to their own affairs. Yet, despite these differences, even aloof gods punish those who behave immorally. For example, in the case of the Greeks, gods' punishment encompassed not only the main sinner, but their descendants as well. In Buddhism good or bad actions affect karmic justice, the causal mechanism that governs the universe. So while sin may not be defiance against a god, it is a transgression against a universal moral code that has dire consequences. Punishment for sin is universal across groups.

Sin is also the basis for ones salvation in this world or the next, if applicable. For Muslims, god simply tallies up good deeds against sins to place ones soul in heavenly bliss or damnation in hellfire. For Christians the acceptance of Jesus is the basis for salvation although leading a sinful life is indication that one has not truly taken him as his lord and savior. There are various religious groups where living a moral life is rewarded with fortune and progeny or protection from harm, others where it leads to reincarnation to a higher life form or even better, release from the cycle of birth and re-birth all together.

Our secular lives are also filled with the concept of moral transgression and judgement. In the absence of belief in supernatural forces, as members of groups, we fear judgement from other members or banishment from the group for our immoral acts. We are afraid of punishment by the law for our misdeeds. Across the world, judicial systems' moral code is heavily influenced by religious values. Of the countries that follow a religious law system, most follow Islamic law (Sharia) or Judaic law (Halakha)in conjunction with other legal systems such as civil or common law, but that is not to say that religious moral values are absent from other legal systems. Even in secular countries like the US with a common law system, Judeo-Christian values can be seen throughout the penal code. But not all sins are similarly important despite ideas such as "all sins are equal in the eyes of God". In practice, homicide as a far more serious sin than telling a lie. That is to say that the concept of moral wrongdoing or sin is universal in our secular lives and a large set of sinful actions that are culturally relevant to our modern lives are also universal. Sin as a universal concept provides an easy avenue to explore system of beliefs or ideologies.

Particularly interesting to this study are the seven deadly sins. There is a long written history and philosophy on these sins and while they are based in Christian doctrine, they provide an interesting case study into sins.

4.3 Origin of the seven deadly sins

The seven deadly sins are a grouping of vices within Christian doctrine. Based on Christian beliefs, *actual* sins can be subdivided into two categories: Deadly sin or mortal sin and forgivable of venial sins. As the name would suggest, a venial sin while creates a barrier to God, it is more minor. Deadly sins on the other hand are intentional barriers that go against God and are serious offenses. More over these sins lead to more sin. Lust for example can lead to adultery. These sins have pre-Christian roots – Greek and Roman – specifically as highlighted by Aristotle [3] in his description of the excellent virtues. In contrast, each virtue in excess can be thought of as a vice. Following in this tradition, a fourth-century monk, Evagrius Ponticus described the eight evil thoughts [73] which later his influential pupil, John Cassian, expounded on his list in the 5th century [24]. Then, in AD 590, Pope Gregory I revised down to the seven deadly sins in his work [63]. Later in the 13th century St. Thomas Aquinas [8] reasoned for the following ordering: pride, greed, lust, envy, gluttony, wrath, and sloth.

Other religions too have similar lists. The seven deadly sins have similarities with the Buddhists' Three Root Poisons (Greed, hatred, aversion) and the Three Pillars of the Dharma (meditation, moral restraint, generosity). Also the Ten Perfections and Ten Fetters as well as the Four Taints, and the Five Strengths and Five Hindrances bare similar ideas.

4.4 Chess rating systems

Today the Elo rating system is used in and outside of chess such competitive multi-player computer games, soccer, American football, basketball, and Major League Baseball. From its start in 1939, the United States Chess Federation (USCF) used a numerical ratings system devised by Kenneth Harkness that tracked each chess player's personal progress. This already was an improvement over simply tracking tournament wins and losses. Yet, in some cases the Harkness system led to inaccurate ratings. Arpad Elo, a master-level chess player at the time, created a rating system with a statistical basis and probabilistic underpinnings, now known as the ELO systems. He describes his system in *The Rating of Chessplayers, Past and Present*, published in 1978. In this model, each player was thought to have a *"true chess"* ability which this system estimated by taking into account the score of each player in a tournament, the expected performance of each performer, and the win or loss outcome. If a novice player wins against a high scoring seasoned player, their score would rise significantly compared to a win against another novice player. The updating to a person's score evolves with the player's class, number of games played, highest score achieved, current rank, and number of expected wins and losses.

The central assumption here is that a player's chess performance in each game is a normally distributed random variable and the mean value of their performance changes slowly over time. So while a player can perform significantly better or worse from one game to the next, their true ability is the mean of their performance. In this sense, chess performance is a latent variable that the Elo rating system attempts to estimate.

Elo's model has some simplifying assumptions that were later improved upon. For example, player's ability has been found to follow a logistic distribution rather then a normal distribution – i.e. weaker players have significantly greater winning chances than Elo's model predicts. Most notably, Mark Glickman who developed the Glicko rating system, extended the Elo model to include a player's "ratings deviation" (RD) or the uncertainty in their a rating. Specifically, a player with a high RD will have a less certain or stable rating that one with a low RD. Players who compete frequently will typically have a lower RD than those who do so sporadically. As such, each player has an interval that captures their true score rather than a number (e.g. 95 percent confidence interval 1250 and 1350). Unlike the Elo system, in the Glicko system, player's updating is not uniform in that player one score increase by x will differ than player two's score decreasing by y. This y is is governed by both players' RD's.

4.5 Study 1: Elo model for incomplete data

It is typical to have a situation where from a set of all possible items, we only have a partial rank ordering from various individuals. Consider an everyday situation where I rank order my favorite to least favorite potato chips as follows:
{Jalapeno, BBQ, Plain (not ruffle), Habanero, Onion Sour Cream, Hone mustard, truffle}

This is an ordering of seven items out of hundreds of potato chip flavors (Lay's alone has over 200 flavors). Someone else can have an ordering of a different set which includes items such as {magic Masala, maple, Jamon, Plain, Durian, BBQ}. If we were to aggregate these incomplete rank ordering, Elo chess rating system is a useful way to do so. This first study shows how.

4.5.1 Methods

The first study shows the success of the Elo system as a methodology. In this study, a diverse pool of participant are placed in two group and each rank order sins from most egregious to least. The first group rank orders the *seven deadly sins* from S_1 while the second group orders eight sins from S_2 –a subset of the seven sins in additional to five new sins. In this case we have sets as follows:

$S_1 = \{Wrath, Greed, Lust, Envy, Gluttony, Sloth, Pride\}$

$S_2 = \{Envy, Wrath, Lust, Theft, Deceit, Incest, Adultery, Abortion\}$

$S_1 \cap S_2 = \{Envy, Wrath, Lust\}$

$S_1 \cup S_2 = \{Wrath, Greed, Lust, Envy, Gluttony, Sloth, Pride, Theft, Deceit, Incest,$
$Adultery, Abortion\}$

Using Elo, a single, stable, group ordering was reached by treating each sin as a *chess player* and each participant's ordering as a *tournament*. If a participant judged a sin to be more egregious than another, the sin would receive a lower rank and be the "winner" of a tournament. By continuously selecting two out of fifteen sins and one participant's ordering, sins' scores are adjusted until a stable rank ordering is reached. Using the most basic clustering algorithm, k-mediods, members of religious groups were found to be similar enough in their responses to be placed together by the algorithm.

4.5.2 Similarity measure between individuals

As a means to measure similarity between two individual's rating system, we utilize Goodman-Kruskal gamma (γ). All pairwise γs are calculated as following:

$$G = \frac{N_s - N_d}{N_s + N_d} \tag{4.1}$$

where N_s are the concordant values (agreements) and N_d is the number of discordant values (disagreements).

The individual with the highest average γ is marked as the group expert.

In addition to γ, we also calculate the Kendall rank correlation coefficient (τ) as a different measure for pairwise concordance. Given ratings $r = (r_1, r_2, ..., r_M)$ and $s = (s_1, s_2, ..., s_M)$ on M items, τ is:

$$d_K(r, s) = \#\{(i, j) : i < j, (r_i - r_j)(s_i - s_j) < 0\} \tag{4.2}$$

In other words this is the number of times r and s order items differently. A rating is just a permutation of items $(1, 2, ..., M)$ and is an ordinal feature. Typical machine learning tools ignore this and just treat ratings as vectors in R^M, or throw out all meaningful "distance" information.

In order to implementing a suite of meaningful machine learning algorithms for complete rank order data we use techniques such as k-medoids. k-medoids is a clustering algorithm similar to k-means or k-medians but it guarantees that the central rating for each cluster belongs to a real data point (i.e., it identifies the "expert" of the cluster). We verify this expert using the highest average γ or τ mentioend above

4.5.3 Chess rating algorithm

In this study, we ask participants to rank moral concepts and use a probabilistic chess rating system, Elo, to determine each group's population-wide rating of concepts. To do so, we treat each moral concept as a chess player, each participant's data as a tournament, and each participant's rating of two concepts against each other determines wins and losses in the tournament.

Chess players (CP) = sins $\in S_1 \cup S_2$
Tournament (T) is a participant's ranking of the sins from worst (1) to least worst (7 or 8 depending on set)
Tournament rank for player A ($T_A R$) is the sin's rank assigned by the participant's ranking of the sins
Win= if in the selected tournament a sin is rated as being worse (i.e. has lower value)
Loss = If in the selected tournament a sin is rated as being less severe (i.e. has higher value)
Global rank for player i ($G_A R$) is the sin's current rank (also R'_A in section 4.5.4).

Specifically, the algorithm selects a T at random, then given the CPs involved in the tournament, selects two CPs, determines win or loss depending on the rank of the CP, and updates both CPs scores. This process is continued until a stable ranking system is achieved. Stability is assessed by measuring minimal rank order change: once the change in Goodman Kruscal gamma of the new rank order at time t to the previous rank at t_0 remains constant for a handful of rounds, the ranking is deemed stable and stops (see 4.1).

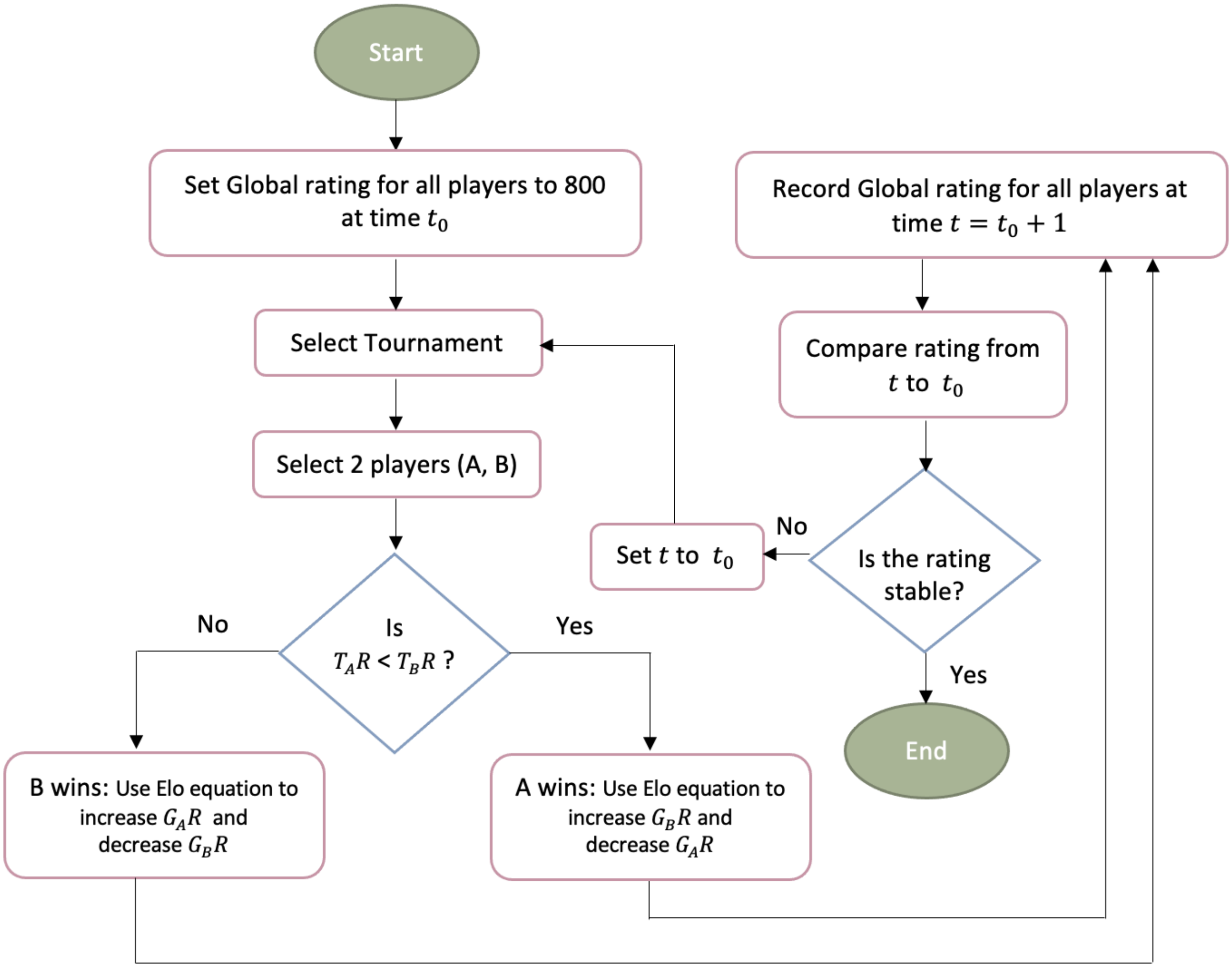

Start
Set Global rating for all players to 800 at time t_0
Select Tournament
Select 2 players (A, B)
Is $T_AR < T_BR$?
No
Yes
B wins: Use Elo equation to increase G_AR and decrease G_BR
A wins: Use Elo equation to increase G_BR and decrease G_AR
Record Global rating for all players at time $t = t_0 + 1$
Compare rating from t to t_0
Is the rating stable?
No
Set t to t_0
Yes
End

Figure 4.1: Elo algorithm for aggregating rank order data of groups.

4.5.4 Elo equations

This implementation deviated from the use of Elo is chess in that there are no draws in this setup as each person is forced to provide a full rank order (i.e. no two sins can have the same rank). Equations are as follows:

$$R'_A = R_A + K(S_A - E_A) \tag{4.3}$$

where

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

R'_A: Player A's new rating

R_A: Player A's old rating

K: a multiplier on updating that affects maximum adjustment

$S_A \in \{0, 0.5, 1\}$: Outcome in a tournament (loss, draw, win)

E_A: Expected score of player A.

With the set up of the game, each player's expected score is calculated by the probability that a CP would be ranked higher or lower as follows:

$$P(A \text{ wins}) = \frac{R_A}{R_A + R_B}$$

The benefit for using the Elo system is that new items can be tested and a group level rating can still be attained if some old items are added to the new item list. This is especially attractive given the large number of moral concepts we would like to study but are limited by participant's inability to rank order large lists of items.

4.5.5 Data

The data was gathered on Amazon Mechanical Turk and from University of California students body with a total sample size of $N = 459$. The predominant groups in the sample were protestant, Catholic, Hindu, Spiritual, Atheist, and Agnostic. With a small number of Muslim, Jews, and Buddhist in the sample, most analyses are done for the larger groups.

Each participant completed the task online.

4.6 Results

4.6.1 Model testing

To determine the effectiveness of the Elo implementation in aggregating each group's ordinal data we: 1. Aggregated only S_1, 2. Aggregated $S_1 \cup S_2$. In order to verify the accuracy of the techniques, we compared the probabilities between the items in $S_1 \cap S_2$ in both the Elo rating output for S_1 and $S_1 \cup S_2$. Specifically, $Envy, Wrath, Lust$ were the mutual items between S_1 and S_2. Based on Luce' Choice axiom [83], the probability of each of these items winning against another should not be different for the two outputs –i.e. "independence from irrelevant alternatives" (IIA). In particular, selection of one of the mutual three items over another in S_1 should not be affected by the presence of additional items such as in S_2. Our findings showed that the $P(Envy, wrath)$, $P(Envy, Lust)$, and $P(Lust, wrath)$ followed Luce's axiom. That is, the probability of selecting one item over another was consistent across the different conditions. This gives us confidence in the model's outcomes.

4.6.2 Group ranking

With the addition of the five new sins, we analyzed the larger groups and derived the following orderings:

	Ad	Ab	In	De	Th	Lu	En	Gr	Sl	Wr	Pr	Gl
Catholic	3	4	1	7	5	12	10	8	11	2	9	6
Mainland	5	7	1	4	2	10	7	6	11	3	9	12
Evangelical	3	4	1	5	8	11	7	6	10	2	12	9
Hindu	4	11	1	3	2	12	7	5	12	6	8	9
Atheist	5	11	2	4	1	8	9	6	7	3	12	10
Agnostic	5	8	3	6	2	12	10	4	9	1	11	7
Spiritual	1	5	12	9	4	7	10	3	6	2	11	8

Table 4.1: Total rank orderings of 12 sins combined. In order from left to right: *Adultery, Abortion, Incest, Deceit, Theft, Lust, Envy, Greed, Sloth, Wrath, Pride, Gluttony.*

The religious groups consistently rank incest as the gravest sin, unlike non-religious groups where wrath, theft, and adultery take the first place. Hindu ranking was the least stable amongst the groups while the Christian groups were the most stable. This points to the lower variability in for religious respondents as compared to the rest.

4.6.3 Clustering results

To maximize the interpretability of the ML technique used, we utilized k-metoids to cluster the data. Clustering was performed on the ranking of S_1 items only. As an ultra-simple example of the effectiveness of this method, consider k=2 clusters.

Typical medoids are:

	Lust	Envy	Greed	Sloth	Wrath	Pride	Gluttony
Cluster 1	7	2	3	4	1	6	5
Cluster 2	1	5	3	4	2	6	7

Table 4.2: Orderings of sins for clusters when number of k = 2.

With the notable difference that the first group considers Lust the least among sins while the other cluster presents it as the worst. With this clustering, 35 percent of one group of surveyed Christians belonged to cluster 1, but 72 percent of Catholics (separate group) did. 93 percent of those identifying as non-affiliated belong to the first group, while 89 percent of Hindus and 79 percent of Protestants are clustered with the second.

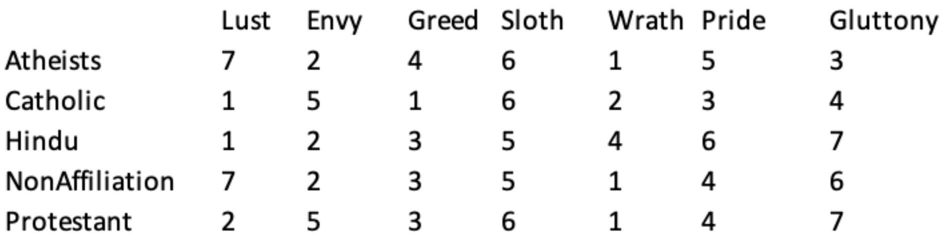
In this case we are identifying a real feature of religious values that matches tendencies in identified religious groups. While the particular orderings for the clusters aren't entirely stable (this is a failing of hard clustering and the ad hoc choice of k) what is consistent is the division of groups along the Lust axis. In fact, 7 and 1 are the most popular ratings. For the other sins, ratings are either generally distributed, or in the case of Wrath, almost always ranked first or second.

With the method checked for the two case scenario, a complete clustering of responses is undertake.

Clustering stability is not absolute (repeated runs of the algorithm produce different clusterings due to dependence on initial random assignment). Though there are some general trends, when N is small, the "Wrath" dimension has very little information. Most people rank it as one of the worst.

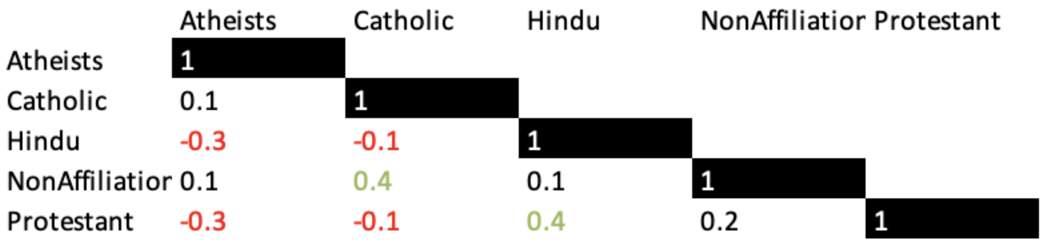
We frequently see a divide among groups that think Greed is or is not among the most severe of the seven.

Some results suggest the method does capture ideological features of the groups. With N=4 clusters, despite the small number of Hindu respondents compared to other groups, one cluster had more Hindus than any other single religious group (and also contained a strong majority 65% of all of the Hindu responses). This group tended to rank greed worse than the other clusters. With N=3 clusters the clearest difference is found in one group whose members tend to rank envy as being very bad.



	Lust	Envy	Greed	Sloth	Wrath	Pride	Gluttony
Atheists	7	2	4	6	1	5	3
Catholic	1	5	1	6	2	3	4
Hindu	1	2	3	5	4	6	7
NonAffiliation	7	2	3	5	1	4	6
Protestant	2	5	3	6	1	4	7

Figure 4.2: Elo output for each group's rating. 1 signifies the worst moral transgression while 7 is the least severe.



	Atheists	Catholic	Hindu	NonAffiliatior	Protestant
Atheists	1				
Catholic	0.1	1			
Hindu	-0.3	-0.1	1		
NonAffiliatior	0.1	0.4	0.1	1	
Protestant	-0.3	-0.1	0.4	0.2	1

Figure 4.3: Each square shows the level of correlation between the groups. Here, Catholics and NonAffiliated as well as Protestants and Hindus are highly correlated.

In this analysis, participants who identified as Spiritual or Agnostic were put into the category of "non-affiliated" as they were small in number. The ranking of each group is shown in 4.2.

The see the level of correlation between groups, we perform a γ calculation between the group ranking of each group. The results show that Catholics and NonAffiliated as well as Protestants and Hindus are highly correlated (see 4.3).

4.7 Study 2: Ranking like a Christian

A person can have multiple group affiliations, be it religious, political, or cultural. These groups form around shared beliefs, activities, and identities. Much of the rituals are to establishing uniform beliefs and behaviors (e.g. prayer, rallies, holidays and festivals). As these beliefs and behaviors are often visible signals to members, non-members also observe them and form opinions. How well do non-members understand religious beliefs?

In the second study using rank order data, we can explore how well known Christian beliefs are. There are three participant groups: i) Atheists are asked to answer on behalf of devout Christians, ii) Christians are asked to answer on behalf of devout Christians, iii) participants (Atheist, Christian, and other) are asked to rank order based on their own judgement. All participants rank order the *seven deadly sins* from worst to least worse.

As far as the State is concerned, Atheism is a religious group and enjoyed the same privileges as other religions –ruling for the Kaufman v. McCaughtry (2005) case. The court held that "atheism is [the plaintiff's] religion, and the group that he leads is religious in nature even though it expressly rejects a belief in a supreme being." Here, the court looked to a number of U.S. Supreme Court precedents where a range of "nonreligious" beliefs were treated as being equivalent to religion:

> "The Supreme Court has said that a religion, for purposes of the First Amendment, is distinct from a 'way of life,' even if that way of life is inspired by philosophical beliefs or other secular concerns. A religion need not be based on a belief in the existence of a supreme being, (or beings, for polytheistic faiths) nor must it be a mainstream faith."

But Atheism as a "religion" for First Amendment purposes does not mean that adherents believe in the supernatural, hold devotional services, or have a sacred Scripture. In fact, the American Atheists clearly states that they are *not* a religion: *"Atheism is not a disbelief in gods or a denial of gods; it is a lack of belief in gods."*. Many Atheists advocate for a complete separation of church and state and are aware of the long history of conflict with religious groups, particularly with Christian groups. An interesting questions we explore is: How well do Atheist know Christians and are able able to accurately answer on behalf of devout Christians? We explore this very question.

4.7.1 Methods

In this study, we ask three groups of participants to rank order the seven deadly sins from worst to least worst as follows:
i) Group 1: Atheists are asked to answer on behalf of devout Christians
ii) Group 2: Christians are asked to answer on behalf of devout Christians
iii) Group 3: participants (that also include Atheists and Christians) answer the questions based on their personal opinion

After computing the aggregate ranking of each group, using γ (see section 4.5.3) we examine the correlation between the three groups. Our expectation is that Atheists in the first group would perform poorly while Christians in the second group will be successful in their task.

4.7.2 Data

The data was gathered in two rounds. First, participants ($N = 721$) participants were asked about their religious affiliations. Second, based on the demographics of participants, a number of self identified Atheists ($N = 42$) and Christians ($N = 54$) completed the task and labeled as group 1 and group 2 respectively. A general public ($N = 460$) which also included Christians and Atheists (among other participants) also completed the task and were labeled as group 3.

Participants were recruited from Amazon Mechanical Turk and the University of California student body and completed the task online.

	Group 1	Group 2	Group 3 Christians	Group 3 Atheists
Group 1	**0**	-	-	-
Group 2	0.13	**0**	-	-
Group 3 Christians	-0.29	0.67	**0**	-
Group 3 Atheists	-0.42	-0.66	-0.2	**0**

Table 4.3: Between group distance.

4.7.3 Results

The task of answering on behalf of a devout Christian was particularly difficult for Atheists (group 1) and quite easy for Christian (group 2) as expected. Atheists were negatively correlated (or far) with how Christians rank ordered the sins ($\gamma = -.29$). To examine whether participants in Group 1 were sincerely attempting the task, in the table below we also show their distance to Atheists who were answering on their own behalf. Even here, participants in Group 1 are further away at $\gamma = -0.42$. We take this an an indication that participants were in fact performing the ranking task to the best of their ability.

Unsurprisingly, Christians in the second group were able to successfully predict what Christians answering on their own behalf would select. Interestingly, the medieval Catholic Church ordering of the sins is uncorrelated with modern day Catholic rankings at $\gamma = 0.04$.

4.8 Discussion

Elo was shows to be an effective tool to aggregate group data and derive insights into groups. In study one, we showed that Elo can successfully put create a single group ranking for each group based on incomplete data. Clustering results and other statistical analyses showed relative differences of how severe sins are to each other. Lust turned out to be a divisive concept while mostly wrath was deemed the worst.

The second study showed how well outside members can predict member responses. Atheists were unable to successfully predict the responses of Christians while Christians answering on behalf of Christians were. Additionally, Atheists were not strongly correlated with each other as a group.

Future studies would benefit from more data and many more items. The seven deadly sins while universal may not have been as equally salient to each group. The variability of Hindu respondents in study one provides some reason to questions how culturally similar these sins are. On the other hand, sins are more commonly associated with expressed actions. The seven deadly sins were states of being, with the exception of wrath which is the only active sin. Unsurprisingly, most individuals identifies wrath as the worst of the group.

Chapter 5

Further evolution of natural categorization systems: A new approach to evolving color concepts

To assess the effectiveness of a simulation-based approach in modeling the evolution of linguistic categorizations, a dynamic model of language evolution is applied to the natural color categorizations of 108 linguistic communities collected in the World Color Survey (Kay, Berlin, Ma, & Merrifeld 1969). Our evolutionary dynamics, modeled after human communication, are specified by the discrimination-similarity and 2-player teacher game (Komarova, Jameson, & Narens, 2007, "Evolutionary models of color categorization based on discrimination," Journal of Mathematical Psychology, 51, *359—382), where color-naming systems are evolved to a stable equilibria through agent interactions. This simulation-based approach remedies the sparseness of empirical, diachronic data, which would be ideal for studying natural evolution trends in real human communities, by broadly approximating these trends in an idealized form. Results suggest that our simulations are a suitable representation of natural evolutionary processes, as evidenced by the fact that all 108 communities' systems were* *evolved to a stable equilibrium while upholding the original integrity of each community's categorizations—that is, not imposing any external structure on the original categorizations. More broadly, we demonstrate that this approach can have valuable insight and real implications for research in fields where diachronic data is sparse. For example, results from these simulations have the capability to evaluate the validity of different hypotheses regarding category evolution. Weighing in on the linguistic debate between the Emergence and Partition Hypotheses, our analyses found evidence in favor of the Emergence Hypothesis. Additionally, this paper presents novel explorations of evaluating equilibrium stability and determining category boundaries.*

5.1 Introduction

A major area of interdisciplinary study investigates how words and signals acquire meaning. This is a central and heavily investigated issue in the humanities (philosophy and linguistics), the social sciences (psychology and anthropology), and engineering and computer science (robotics and artificial intelligence). This article investigates a special case: the evolution of color concepts in languages of non-industrialized, isolated linguistic communities.

Color naming has a long history in academia, beginning with seminal work on ancient Greek color terminology by Gladstone in 1858 [**?**] that was extended by other 19th century researchers to additional ancient languages. The subject gained increasing recognition in 1969 due to the seminal study of Berlin & Kay's, *Basic Color Terms: Their Universality and Evolution* [16], and later from the *World Color Survey* (WCS) by Kay, Berlin, Maffi, & Merrifeld [68]. The latter surveyed the color naming and categorization behaviors of over 2500 individuals from 110 non-industrialized and isolated linguistic communities world-wide. The WCS consisted of two naming tasks completed on a standardized set of 330 unique color chips. This produced a famous data set that has been analyzed from a number of perspectives using a variety of methodologies. This article provides a new methodology for analyzing data from the WCS based on evolving naming strategies through evolutionary game theory.

5.1.1 Rationale for Game Theoretic Approach to World Color Survey

There is a history of using game theory to understand the evolution of meaning in linguistic and signaling systems. There are many approaches to this subject (e.g., [120, 121] [117] [50] [23] [**?**] [19] [18]), and several have been used for color naming.

We follow this path and consider color names as conventions [116], and use the *discrimination-similarity game* and the *2-player teacher game*, developed in Komarova, Jameson, & Narens [72] as a way to evolve population color naming strategies. The dynamics of these games is based on a form of reinforcement learning. Unlike some other evolutionary color naming models that emphasize supposed properties of human color vision (e.g. the Hering primaries of red, blue, green, and yellow lights having special perceptual properties and salience), the game dynamics are based only on primitive ideas about communication and the agents' ability to discriminate one colored stimulus from another.

The WCS revealed a highly restricted pattern of naming strategies across linguistic communities [45]. A number of theories have been developed to explain this pattern. The idea of Berlin & Kay and others, who prescribe to the universalist view, is that each community is in a particular stage of color naming evolution. The overall evolutionary process of cultural color naming can be understood by analyzing the increasing complexities of the naming strategies in terms of how populations partition the color stimulus space into a set of concepts called “basic color terms” (BCTs).

Our analysis and goal for the WCS data is different. Like Berlin & Kay, we take each linguistic community to be in a particular stage of color category evolution, but we do not then use another unrelated language with an additional BCT to predict the next, more complex stage of that linguistic community evolution. Instead, we ask how well evolved, *as a communication system*, is the community's naming strategy, and what would its future evolution look like if its population communicated freely about color? We model this in our evolutionary dynamics by using individual data from the WCS to model agents in a society and having them play the communication game repeatedly until an equilibrium strategy (i.e. a color naming convention) arises and remains stable.

This evolutionary approach is particularly useful for two reasons: First, much of the data available on color naming, including the WCS, is cross-sectional. To investigate the evolution of naming systems, diachronic data would be most ideal because it would provide real observations of how a community's naming system changes over time. However, data of this type is extremely sparse, and now impossible to attain for the linguistic communities that have gone extinct or are composed of bilingual speakers. Therefore, simulations can generate psuedo-diachronic data through repeated agent interaction modeled after human communication. Evolving the cross-sectional data in this way provides a unique perspective on possible ways which color naming systems evolve, which has not been done previously.

Second, if the evolutionary dynamic proves successful, it opens up more possible avenues of exploration. For example, the pioneering work of Berlin & Kay [16] grouped languages into evolutionary stages based on the number of BCTs they contained. They applied their linguistic approach to determine the number of BCTs in the 110 languages collected in the WCS. Fider et al. [29] confirmed many of the conclusions made with regards to the WCS, but instead determined the number of BCTs using a computational method that requires no knowledge of semantics of a language's color words. However, both these methods do not provide information on where languages lie within a stage—that is, did a language recently enter a stage by acquiring a new BCT? Or is a non-basic term growing in prominence within a language, positioning it on the cusp of entering the next stage? By evolving a single linguistic group's data independently to a stable equilibrium, we can *(i)* observe the evolution of individual color concepts and draw conclusions about terms that are "falling in" or "out" of basicness, and *(ii)* propose a new grouping criterion for comparison across naming systems that provides a richer understanding of a language's evolutionary path.

5.1.2 Theories of Color Categorization

Berlin & Kay modeled their theory so that every color always possessed a unique color name (i.e. there existed no uncertainty or unnamed colors). This is reflected in the data they collected for their 1969 book [16] and in the World Color Survey [68]. Each participant was required to assign a name to each stimulus in the color space. This allowed them to partition the color space into color categories and identify which of these were considered "basic". However, some argued and presented data (e.g. Levinson 2000 [74]), that it was possible to have gaps in conceptual color space. In other words, there could exist regions in the color space for which there are no color concepts or names. This gives rise to two different theories of color term evolution: *Partition* and *Emergence Hypotheses.*

In the Berlin & Kay theory, the introduction of a new term can only arise by splitting an existing concept because the whole space is named. Japanese provides a good example of this with the single term *aoi* for blue and green hues that split in the 14th and 15th century A.D. into *aoi* for blue and *midori* for green. It also happened in Russian with the splitting of its blue category—originally called *sinij*—into two categories, *sinij* for darker blues and *goluboy* for lighter blues. This hypothesis of color category evolution is called the *Partition Hypothesis.*

Levinson [74] proposed a different hypothesis, one where in the earliest stages of color evolution, color terminology did not partition the entire color space, but rather had ambiguous, unnamed regions. He supported his claim with the data he collected on the Yélî Dnye people. He observed that color terms initially remain focused around certain substances or objects and these terms became more generalized over time, expanding further into the color space. This theory proposed that whenever the need arose for new color terms, they were introduced to fill in the unnamed regions. This theory is called the *Emergence Hypothesis.*

Both theories hypothesize about emergence of color categories, space partitioning, and color terms. Given the expansive lexicon of color terms contained in many languages, Berlin & Kay achieved a more targeted analysis through their formalization of basic color terms.

5.1.3 Basic Color Terms & the World Color Survey

This idea of "basic color terms" (BCTs) originates from the *universalist* perspective of the *linguistic relativity* debate. Universalism states that color cognition is a universal, physiologically based phenomenon much more so than a cultural one. Berlin & Kay popularized the universalist view by exploring universal features of color categorization through a formalization of BCTs introduced in their book. They defined the set of BCTs as the smallest set of color terms within a language with which a speaker could name every possible color. For example, Berlin and Kay identify 11 BCTs in English: black, white, red, blue, green, yellow, orange, purple, pink, brown, and gray. Their theory suggests that for all languages there exists: *(i)* universal constraints on possible number of BCTs, and *(ii)* a universal pattern of evolution for the emergence of new BCTs. Data used to support these claims were collected from "published sources and personal communication with linguists and ethnographers who have specialized knowledge of the languages in question" on 98 language groups and survey data from 20 language groups in the San Francisco Bay Area.

However, this empirical data was met with criticism due to the sparse number of participants per language (as few as one participant per language), bilingual nature of participants (who spoke English in addition to the target language), and location of data collection (San Francisco rather than the homeland of the target languages). In response to such concerns, Berlin & Kay initiated a multinational survey referred to as the World Color Survey.

The *World Color Survey* [67] which was undertaken as an effort to verify the claims made in Berlin & Kay's 1969 book [16], gathered data on the color categorization of 110 monolingual tribes around the world, each with ∼24 participants on average. The data [22] was gathered in person by linguistic missionaries who conducted two distinct tasks: a mapping task and a naming task. In the mapping task, participants were given a set of color words and were asked to identify which color chips from the 330 standardized Munsell stimulus set best represented each color word. Each participant's chosen set of chips are his/her "focal chips". In the naming task, participants named 330 color chips that were presented on a gray background, one at a time, in a fixed, random order.

The data from these two tasks, taken from 108 of the languages included in this survey– Language numbers 62 and 93 were omitted from analysis in this paper because of the poor nature of the data collection at the time of the survey. Any future mentions to "all WCS language" refers to the set of languages that excludes languages 63 and 93–are the basis for our new methodology to study the evolution of color categorization.

5.2 ColorSims: An Evolutionary Game Theoretic Tool

In this framework, naming strategies of simulated populations evolve to stationary equilibria as the agents play a simple communication game. A stationary equilibria is defined as a non-Nash stable system that undergoes minimal change over a long period of time. The communication game requires agents to assign names to color stimuli and "communicate" repeatedly with members of the population until a naming convention arises. The agents in these evolutionary dynamics are endowed with minimal perceptual and learning abilities. A Python-based program written by S. Tauber, called *Colorsims 2.0* [39], was developed to model the communication game specified in Komorova et al. [72]. This platform has been utilized in several past studies to evaluate the evolution of color naming systems in populations of simulated agents on a one-dimensional color space [71, 72, 60, 60, 96].

The simulation framework used in this paper is an extension of *ColorSims* with *(i)* a higher dimensional color space that more realistically approximates human color perception, *(ii)* real observer population data taken from the WCS, *(iii)* a measure to assess the stability of a color naming convention. This updated version of the software is referred to as *ColorSims 2.0* [39].

5.2.1 Games

The evolutionary dynamics are comprised of two steps. First, agents discriminate between like and unlike stimuli by assigning names to the stimuli. Second, agents' naming strategies are either updated or learned from another agent's strategy depending on the agreement of their assigned names.

Discrimination-Similarity Game

The *discrimination-similarity game* was introduced by Komarova et al. [72] as a mechanism for populations of artificial agents to create shared categorization systems. In the game, agents make judgments about how "close" or "far" two stimuli seem from each other in their color appearance. In order for these judgments to be comparable across agents, Komarova et al. defined a measure *k-sim* which serves as an objective criterion for determining whether two colors are "close" or "far" in appearance. Komarova et al. (2007) write,

> $k_{sim}(a, b)$ is interpreted as being related to the utility of categorizing a and b as the same or different colors. It is defined by the environment and the life-styles of the individual agents. It is used to reflect the notion of the pragmatic color similarity of the patches. For instance, suppose one individual shows another a fruit and asks her to bring another fruit "of the same color." It is a nearly impossible task to bring a fruit of a color perceptually identical to the first, because different lighting, different color background and slight differences in fruits' ripeness contribute to differentiating its perceived color from the comparison fruit. Therefore to satisfy "of the same color" of a fruit's ripeness in practical terms, the individual must be able to ignore such unimportant perceptual differences and bring a fruit that is "of the same color" practically. It may also be as important to be able to distinguish ripe, edible, "red" fruit from the unripe, "green" ones.

At the beginning of each round of play, two agents—Player 1 and Player 2—are randomly chosen from the population. Both players are presented with an independently and randomly selected pair of colored chips and are individually asked to provide a name for each chip. Players assign names to a chip based on their probability strategy matrix (see Section 5.2.3). Two chips should be considered of the same category and given the same color name if the two chips are within 1 *k-sim* of each other. Conversely, if their distance is greater than 1 *k-sim*, they should be considered of distinct categories and assigned different color names. A player is awarded a "personal success" if the assigned names of the two chips are consistent with the criteria above, and is awarded a "personal failure" if the criteria is not met. The round is considered a "social success" if and only if both players have personal successes and both assign the same names to the same chips, otherwise it is considered a"social failure". More specifically:

1. *Social success*: *Personal success* for Player 1, *Personal success* for Player 2 (e.g. Chip a and chip b are outside 1 k-sim, Player 1 chose α for chip a and β for chip b, Player

2 chose α for chip a and β for chip b)

2. *Social failure*: *Personal success* for Player 1, *Personal success* for Player 2 (e.g. Chip a and chip b are outside 1 k-sim, Player 1 chose α for chip a and β for chip b, Player 2 chose θ for chip a and γ for chip b)

3. *Social failure*: *Personal success* for Player i, *Personal failure* for Player j

4. *Social failure*: *Personal failure* for Player 1, *Personal failure* for Player 2

In the case of social success and social failure resulting from two personal failures—cases 1 and 4—players perform a form of reinforcement learning [72] by updating their probability matrices accordingly: in case 1, players strengthen the probability of choosing α and β when naming chips a and b, respectively; in case 4, players weaken the probability of assigning α/θ and β/γ to chips a and b, respectively. This concludes a round of the *discrimination-similarity game*. For all other cases, players continue to another game, called the *2-player Teacher Game*.

2-player Teacher Game

The *2-player teacher game* was also introduced by Komarova et al. [72] as a particular implementation of a reinforcement learning dynamic. When the outcome of the *discrimination-similarity game* results in social failure but at least one personal success for a player—cases 2 and 3 (see Section 2.1.1)—the same players from the previous game will further engage in the *2-player teacher game*. The game appoints the player with the personal success as the *teacher* (in the case of two personal successes, one will be chosen at random) and the remaining player as the learner.

In case 2, one player is chosen at random as teacher, say Player 1, and Player 2 as learner. Player 2 strengthen the probability of choosing α and β when naming chips a and b, respec-

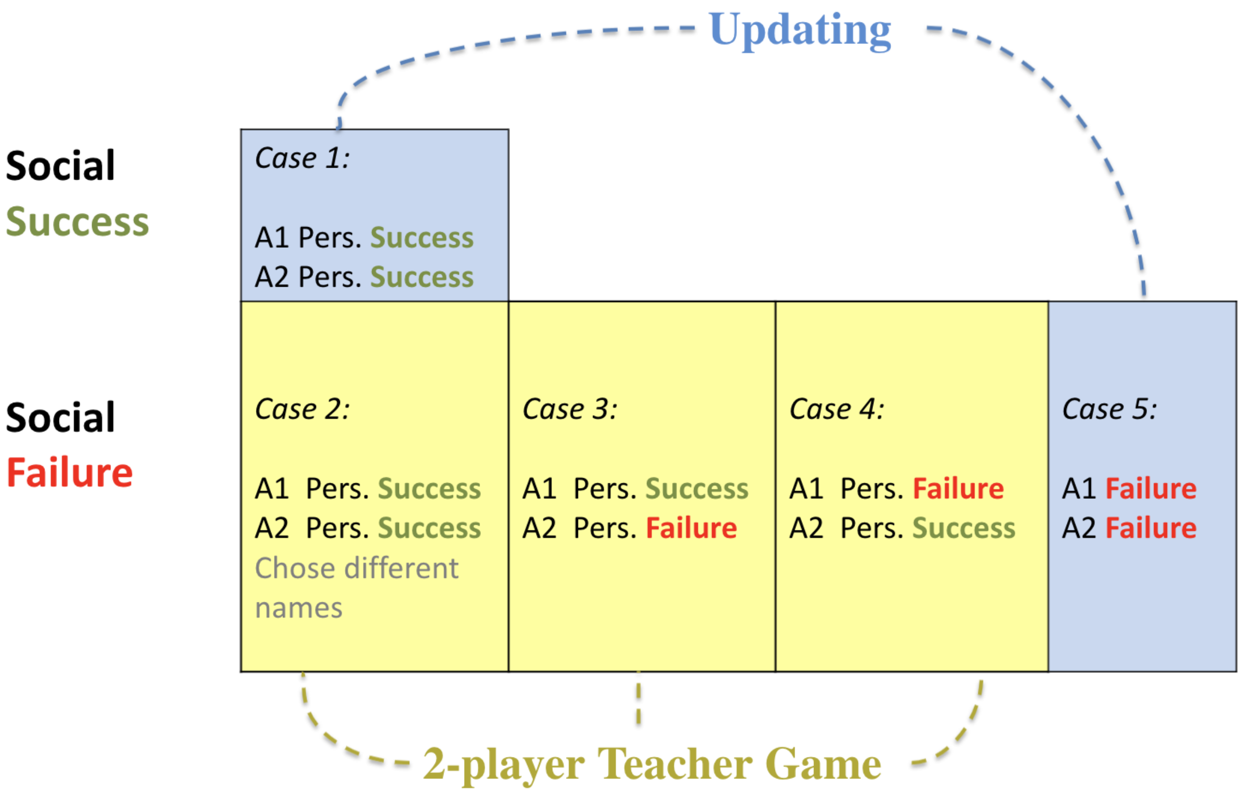


Figure 5.1: Decision to update or go to two-player teacher game.

tively and weakens probability of choosing θ and γ. The teacher simply strengthens their probabilities for choosing α and β. In case 3, the player with personal success is assigned as the *teacher* and the other player as the learner. The learner strengthens its probabilities based on the *teacher's* assigned names and weaken their chosen probabilities. The teacher, similar to case 2, simply strengthens its own strategies as follows:

Social success:

(i) Personal success for Player 1 and Player 2

Given that there is social success, it must follow that $\alpha = \mu$ and $\beta = \nu$. Hence, Player 1 and Player 2 update as follows:

$$P^1_{i,\alpha} \rightarrow P^1_{i,\alpha} + q, P^1_{j,\beta} \rightarrow P^1_{j,\beta} + q$$

$$P^1_{i,\sigma} \rightarrow P^1_{i,\sigma} - q, P^1_{j,\tau} \rightarrow P^1_{j,\tau} - q$$

$$P^2_{i,\alpha} \rightarrow P^2_{i,\alpha} + q, P^2_{j,\beta} \rightarrow P^2_{j,\beta} + q$$

$$P^2_{i,\sigma} \rightarrow P^1_{i,\sigma} - q, P^2_{j,\tau} \rightarrow P^2_{j,\tau} - q$$

where q is an arbitrary probability and σ and τ are two independently, randomly selected chips other than α and β.

Social failure:

(ii) *Personal success for Player 1 and Player 2* Then a random player, say Player 1, is chosen to be the *teacher* and the two players update their strategies as follows:

$$P^1_{i,\alpha} \rightarrow P^1_{i,\alpha} + q, P^1_{j,\beta} \rightarrow P^1_{j,\beta} + q$$

$$P^1_{i,\sigma} \rightarrow P^1_{i,\sigma} - q, P^1_{j,\tau} \rightarrow P^1_{j,\tau} - q$$

Player 2 learns from Player 1's naming schema and updates as follows:

$$P^2_{i,\mu} \rightarrow P^2_{i,\mu} - q, P^2_{j,\nu} \rightarrow P^2_{j,\nu} - q$$

$$P^2_{i,\alpha} \rightarrow P^2_{i,\alpha} + q, P^2_{j,\beta} \rightarrow P^2_{j,\beta} + q$$

(iii) *Personal success for Player i, Personal failure for Player j*

Player i takes the role of the *teacher* and Player j the learner. Player j updates according to the *teacher's* probabilities as in case (ii)

(iv) *Personal failure for Player 1 and Player 2*

Player 1 updates as follows:

$$P^1_{i,\alpha} \to P^1_{i,\alpha} - q, P^1_{j,\beta} \to P^1_{j,\beta} - q$$

$$P^1_{i,\sigma} \to P^1_{i,\sigma} + q, P^1_{j,\tau} \to P^1_{j,\tau} + q$$

where q is an arbitrary probability and σ and τ are two independently, randomly selected chips other than α and β.

Similarly, Player 2 updates as follows:

$$P^2_{i,\mu} \to P^2_{i,\mu} - q, P^2_{j,\nu} \to P^2_{j,\nu} - q$$

$$P^2_{i,\sigma} \to P^2_{i,\sigma} + q, P^2_{j,\tau} \to P^2_{j,\tau} + q$$

The games are repeated until a stable, population-wide naming convention is reached. This concludes a round of the *2-player Teacher Game*.

5.2.2 Evolutionary Dynamics

The evolutionary dynamics arise through repeated rounds of the *discrimination-similarity game* and the *2-player teacher game*. The two agents and two stimuli used in each round are randomly selected before the round starts. Therefore, a simulation with r rounds and N agents will result in $\frac{r}{N}$ interactions per agent on average. At the start of the simulation, social failures are high in frequency. As players engage in more interactions and update their naming strategies, a population-wide naming system begins to emerge as the level of agreement in chosen category names increases (see Figure 2). In other words, the population begins to partition the stimulus space similarly. If a population has reached an optimal, stable naming convention and continues to maintain that stability, we say that the dynamics has reached a solution.

In this paper, a solution is considered to have reached convergence only when agent error is minimized and the optimal number of categories is reached. When optimality is obtained, all color categories are of roughly equal size. The solution to the evolutionary dynamics is always a non-Nash equilibrium: although error is minimized across the stimulus space, errors will persist at the category boundaries. This occurs because, at the boundaries, there always exists two chips, a and b, that are within *k-sim* that should both be named α but happen to belong to different categories and are consequently given different names. Such a scenario continuously leads to personal and, therefore, social failures. Such errors do not influence the overall stability of the population-wide naming convention, though, as the certainty of names for non-border chips remains high. Hence, once the dynamics emerges on a solution, it is stochastically stable as the population-wide naming system experiences minimal change over a very long period of time (see Figure 5.3).

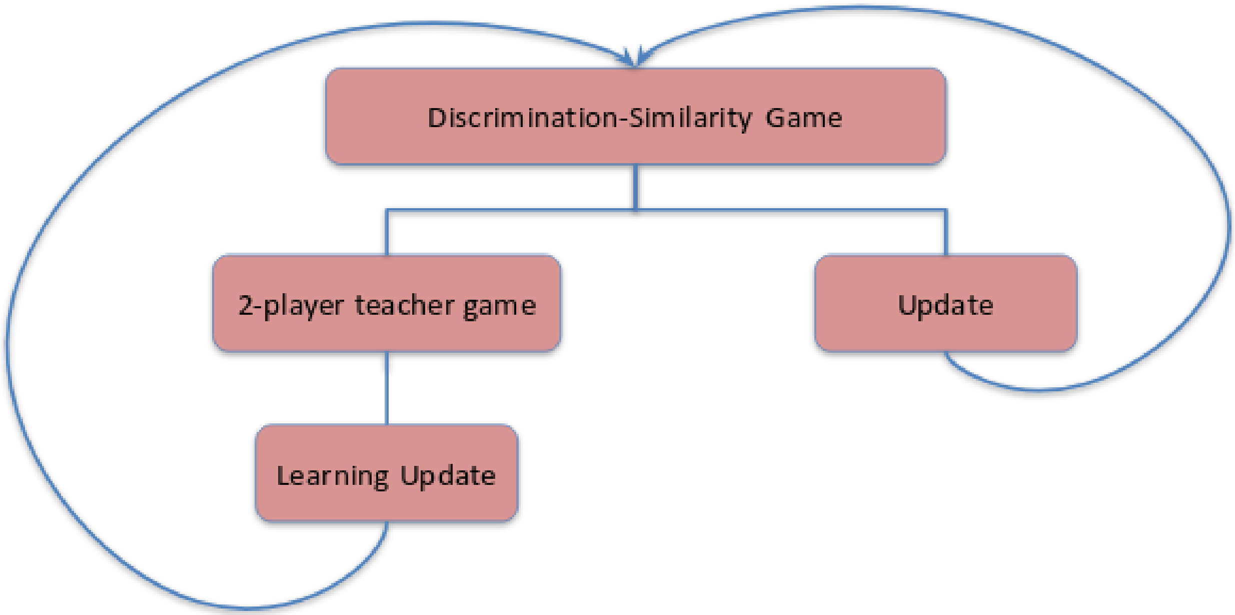


Figure 5.2: Flowchart of evolutionary dynamics.

Evaluating Solution Stability

There are two measures used to determine the stability of a population's color naming system:

1. The level of lexicon agreement across the whole agent population
2. The amount of change in the global agreement level across various time periods

Global lexicon agreement on round r of the game (A_r) is defined as follows:

$$A_r = \sum_{i=1}^{320} \frac{P_{i,\alpha}}{N} \tag{5.1}$$

where r is the game round number, i is the chip number, α is the most frequently used name for chip i, N is population size, and $P_{i,\alpha}$ is number of agents who assign name α to chip i.

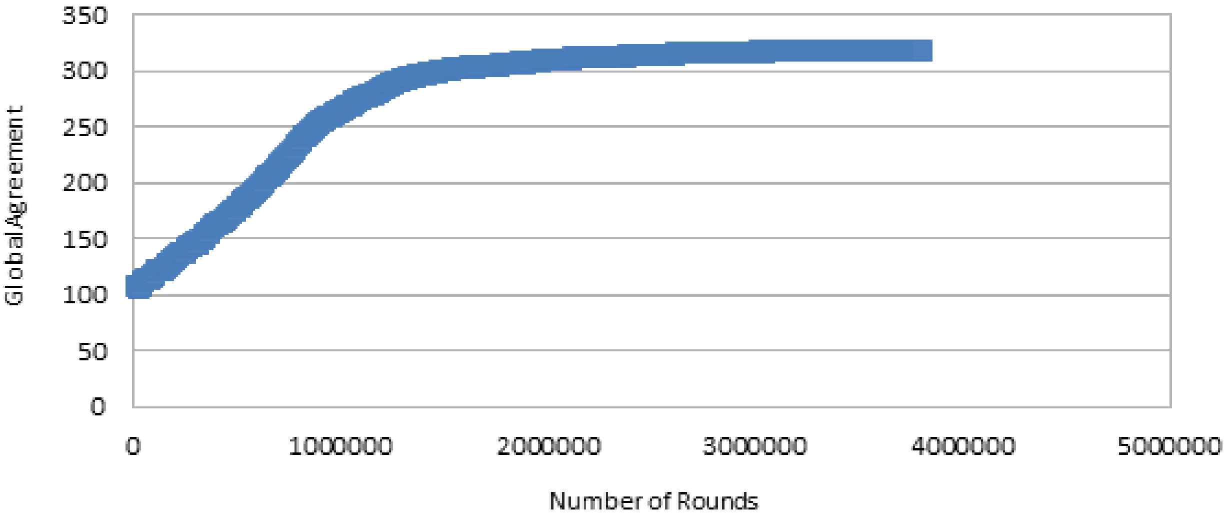


Figure 5.3: Plot of global agreement agreement measure (A_r) of a population of random agents across the number of rounds played.

The global agreement level A_r is designed to be a measure of the strength of agreement between individual agents' color naming strategies within a population. A high A_r indicates large consensus in the population over their categorizations and low A_r indicates low consensus, or a large amount of variability between different agents' individual strategies.

The global agreement level is then used to calculate the *Stability Measure*, defined as follows:

$$S_r = \frac{A_r - A_{r-10,000}}{A_r + A_{r-10,000}} \tag{5.2}$$

Global agreement (A_r) is calculated every 10,000 rounds, so S_r is interpreted as the change in agreement level between rounds r and $(r - 10,000)$, or in other words, the change in agreement level between each "snapshot" of the population.

The simulation marks an r^*, which is the round number at which the stability measure first falls within some predetermined "stability range" (default range is (-0.00175, 0.00175),

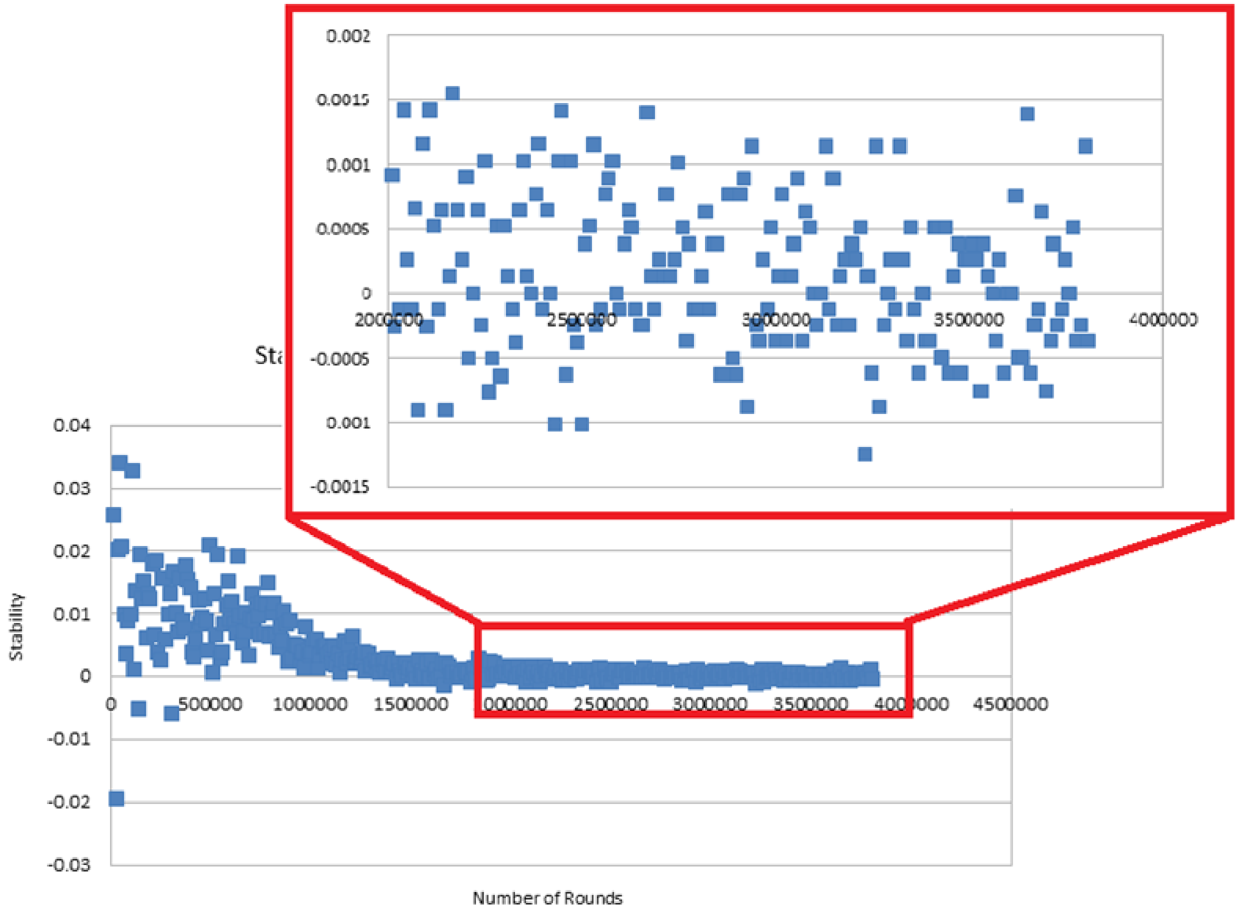


Figure 5.4: Plot of the stability measure (S_r) across numbers of rounds played. We define a "stable" solution to be one which falls within some "stability range" for many rounds. Due to the confusions that agents will have at the boundaries of two color words, the solution will be perpetually subject to minor fluctuations and changes.

determined empirically). The simulation will stop if the solution stays stable—S_r stays within the stability range—for another r^* rounds. That is, a naming system is considered to have converged to a stable solution if the system is stable for r^* rounds, resulting in an overall total of $2r^*$ simulation rounds.

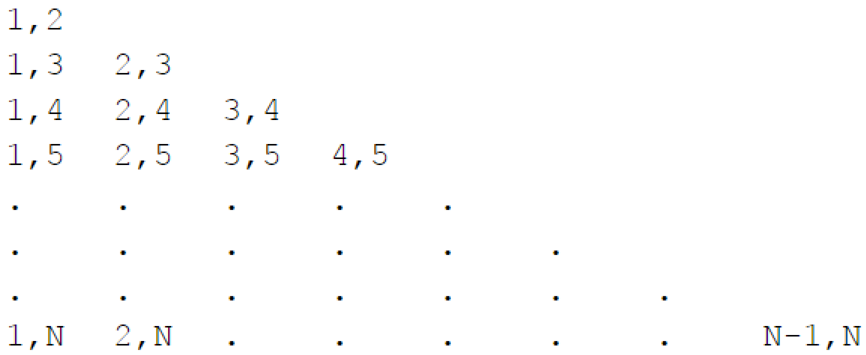
We define S_r in order to understand at which round r the population has converged on a naming system. In the original *ColorSims* program, the number of rounds was a parameter given to the simulation. For simulations with "large" parameters, such as populations with many agents or stimulus sets with many chips, determining how many rounds to run the simulation was an educated guess based on trial and error. Naming systems at the end of those predetermined number of rounds were not guaranteed to be stable. Therefore, S_r ensures two things: (i) the system will run for as long as it needs to reach convergence, as determined from S_r falling within a stability range, and (ii) we can check that the solution remains stable over a long period of time, as determined by r^*.

Defining S_r in this manner ensures two crucial properties: *(i)* the system will run for as long as it needs to reach convergence, as determined from S_r falling within a stability range, and *(ii)* we can check that the solution remains stable over a sufficiently long period of time, as determined by r^*.

For robustness, an alternate measure of global measure was defined called C_r. This alternate measure of agreement compares naming strategies between pairs of agents whereas A_r calculates agreement based by chip. These measures have been found to be highly correlated (r = 0.91). Therefore, for simplicity we use A_r exclusively. The pairwise agent comparison measure (C_r) is defined to be an alternative method for evaluating the level of agreement of a population's color categorization at a given time. Where as the global agreement measure (A_r) aggregates agreement across the set of color chips (see *Section 2.3*), pairwise agent comparison aggregates agreement across all unique pairs of agents. C_r is formally defined as follows:

$$C_r = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=1}^{320} [W_{ik} = W_{jk}]}{\frac{(N-1)N}{2}} \tag{5.3}$$

where i and j are different agents from the population, N is the population size, k is the chip number, r is the number of rounds, W_{ik} is the word that agent i assigned to color chip k, the number of possible pairs of agents is given by $\frac{(N-1)N}{2}$ (see Figure 7), and $[W_{ik} = W_{jk}] = \begin{cases} 1 & \text{if } W_{ik} = W_{jk} \\ 0 & \text{if } W_{ik} \neq W_{jk} \end{cases}$.



```
1,2
1,3  2,3
1,4  2,4  3,4
1,5  2,5  3,5  4,5
.    .    .    .    .
.    .    .    .    .    .
.    .    .    .    .    .    .
1,N  2,N  .    .    .    .    .    N-1,N
```

Figure 5.5: All possible pairs of agents used in calculating pairwise agent comparison (C_r).

5.2.3 ColorSims 2.0 Initialization

Agents

In the dynamics, every agent is endowed with a probability matrix that is updated at the end of each round of interaction. Agents were originally initialized with random probabilities as a basis to test the model and its parameters. Subsequent iterations of the simulation framework initialized agents with real observer data from the WCS.

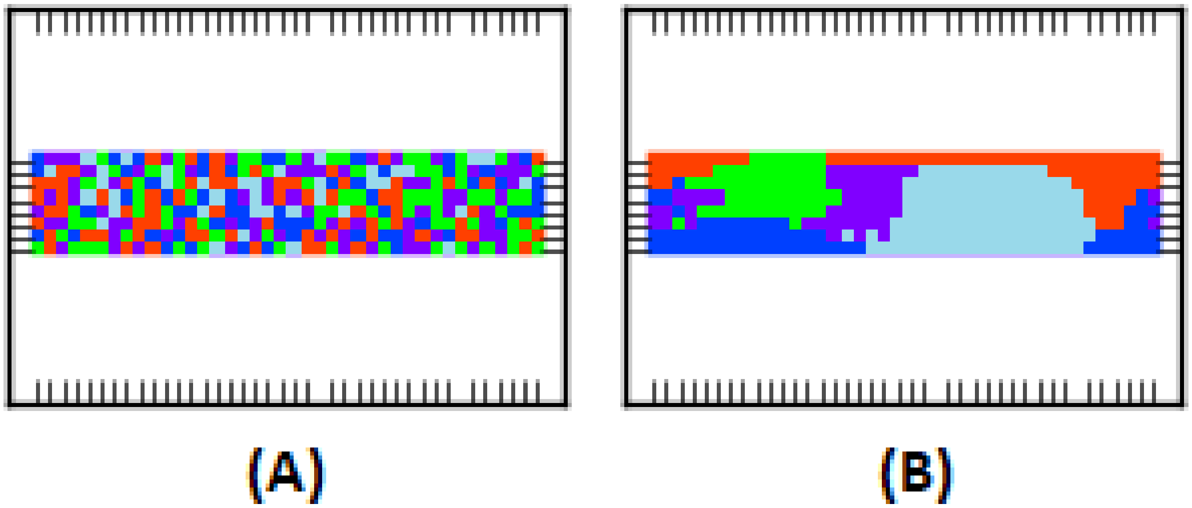


Figure 5.6: Visual representation of the color naming system of a single agent in a random population at the (A) beginning of the simulation and (B) end of the simulation. The colored rectangle represents the two-dimensional set of stimuli, where each pixel in the rectangle represents a single stimulus chip. The color of each box represents the name which the agent is most likely to assign to that color chip.

(i) Initializing Random Agents

Previous work using the ColorSims framework used populations of "random" agents. An agent's probability matrix is organized with colored stimuli along the columns and all possible color names along the rows (see Figure 4, A). A random agent's matrix is populated with random fractions such that column j of the matrix defines a probability distribution for chip j over all possible color terms. Therefore, the value in cell (i, j) represents the probability that this agent would assign the name i to color chip j. We employ the same process to initialize random agents in our ColorSims 2.0 framework.

When an agent is asked to provide a name for chip j, the name that the agent assigns to chip j is the result of a probabilistic draw over the agent's vocabulary according to the distribution defined in column j of the probability matrix.

Though these random agents are not representative of real observers, they are useful because they allow the evolution of the color naming systems to proceed *de novo*. By initializing random agents, we impose no restrictions on the initial state of the system categorization. Since agents can reach a shared, stable naming system with minimal assumptions, we can then evaluate how a system that is presumed to still be evolving, such as those from the WCS, might proceed to stability.

(ii) Initializing Agents with WCS Data

In our simulation-based investigation of color categorization, we initialized agent populations with the WCS naming-task data of each language group. Each agent represents one of the WCS participants. For example, language groups with 25 participants are represented by 25 agents. A cell (i, j) in an agent's probability matrix was given a value of 1 if the participant named chip j with name i or was given a value of 0 otherwise (see Figure 5, A). The agents' probability matrices were updated with repeated interaction between agents, characterized by the dynamics detailed in Section 5.2.1.

Higher-dimensional Color Space

The set of stimuli used in these simulations is a two-dimensional array of colored stimuli. Each individual stimulus in the array is referred to as a color chip. This array is a subset of the standardized set of 330 (320 chromatic, 10 achromatic) Munsell color chips, which is derived from a Mercator projection of a three-dimensional Munsell Book of Color perceptual color space. The set of stimuli used in the simulations is identical to the chromatic component of the stimulus palette used in the World Color Survey [106]. Chips in the grid are organized along the rows according to eight Munsell brightness (value) values and are organized along the columns according to the forty Munsell hue values.

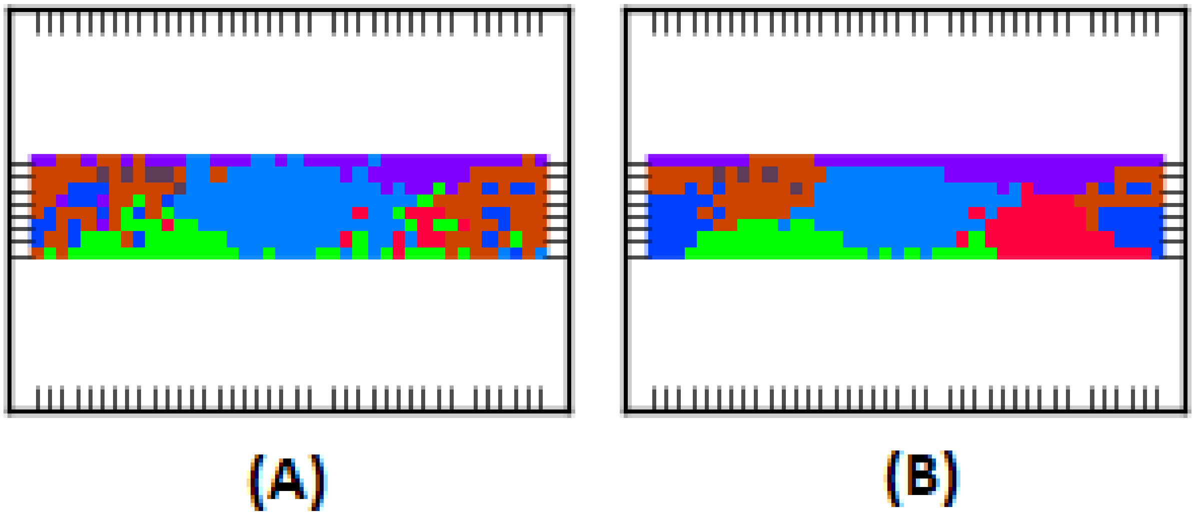


Figure 5.7: Visual representation of the color naming system of a single participant from WCS Language 16 at the (A) beginning of the simulation and (B) end of the simulation. The naming strategy in (A) represents the actual names that this participant assigned to each of the chips in the grid at the time of the World Color Survey.

For the purposes of employing a standard color difference metric, each of the chips in this color grid is mapped to its corresponding coordinates in the three-dimensional CIELUV color space, which in essence allows for a principled assessment of agents' color discrimination judgments based on a standardized uniform perceptual color space. CIELUV was developed by the International Commission on Illumination (CIE) in 1976 with the aim of creating a perceptually uniform space. When presented such stimuli in our evolutionary game, agents ostensibly "perceive the colors" as 3-tuples (L^*, u^*, v^*), the coordinates of the colored stimuli in CIELUV space. While the set of stimuli is constricted to a two-dimensional grid, the underlying space that agents perceive and judge is represented as three-dimensional.

Whereas previous iterations of these simulations utilized a one-dimensional hue space, called a hue circle (i.e. a discrete array of color chips arranged according to just-noticeable-differences) [39], the simulation framework used in this article, implements a three-dimensional perceptual color space in order to create a more realistic representation of how humans perceive color. Most human beings have genes that allow the expression of three retinal photopigment classes–S, M, L. The S-cone peaks in the blue region at 420–440 nm, M in the green region at 534–555 nm, and L in the red region at 564–580 nm. Thus, in principle, all human color sensation can be described as a function of three parameters corresponding to levels of stimulus from these three kinds of cone cells. This biological fact serves as motivation for representing color in three-dimensional models.

The CIELUV color model is appropriate for use here since its approximate perceptual metric permits computation of color difference, or Delta-E, by employing the fact that Euclidean distances between colors map to similar distances in our perception. Therefore, by using a perceptually uniform space, agents can use Euclidean distances between chips in CIELUV as a basis for judgments of color discrimination when engaged in naming game scenarios used to evolve the population's naming conventions.

Color Discrimination Measure: k-sim

The categorizing algorithms in this article are based on the following idea: colors that are perceptually similar to one another are highly likely to belong to the same category. This idea is driven by the following principles: (i) need for categorization is important; (ii) useful categorization should aim to minimize ambiguity, and (iii) when color is a salient signal for categorization, it is more likely that objects which appear similar in color will be categorized together than objects that appear dissimilar in color. These three principles are summarized by the similarity measure, *k-sim* [72].

k-sim is defined to be the minimum distance at which it becomes important, for pragmatic (and not perceptual) purposes, to treat two color chips as belonging to different color categories. Given the pragmatic importance of categorization (principle i), two chips that are within *k-sim* should be are considered to be of the same category (principle ii) and two chips that are outside *k-sim* should be considered to be of different categories (principle iii).

When the stimulus set is one-dimensional (i.e. hue circle), distance between color chips i and j is defined as the number of physical chips that are between i and j. However, when the stimulus set is two-dimensional, we define the distance between color chips to be their perceptual distance. Since the CIELUV space aims to be perceptually uniform, calculating color difference as distances in the physical space will in essence map to similar distances in human perception. Thus, distances between stimuli in the three-dimensional color space are calculated using the standard Euclidean distance metric:

$$d = \sqrt{(L_1^* - L_2^*)^2 + (u_1^* - u_2^*)^2 + (v_1^* - v_2^*)^2} \tag{5.4}$$

where d = distance between two chips, (L_i^*, u_i^*, v_i^*) = coordinates of chip i in CIELUV space.

For simulations using a one-dimensional color space, Komarova et al. [72] developed a mathematical formulation relating the optimal number of color categories to the total number of chips in the stimulus set and the length of *k-sim* to,

$$C^* = \frac{Q}{\sqrt{2k_{sim}(k_{sim}+1)}} \tag{5.5}$$

where C^* is the optimal number of terms and Q is the number of chips in the stimulus set. This formula can be used to derive the appropriate *k-sim* to use when initializing *ColorSims 2.0* simulations by setting $Q = 320$ and C^* equal to the number of BCTs identified by Berlin and Kay. Using these values, we can then solve equation 5.5 accordingly to find *k-sim*. When this formula was developed, the color space was assumed to be one-dimensional and consequently *k-sim* was defined as a distance according to just-noticeable-differences (*jnd*s). Though we are now utilizing multi-dimensional color spaces and measuring *k-sim* according to perceptual distances instead of *jnd*s, we justify using this method to find *k-sim* because, regardless of its underlying metric, k-sim continues to captures the same thing—how close or far colors are perceptually.

5.3 Results

Results are organized into two related but distinct sections: *(i)* a thorough investigation of the validity of our approach and *(ii)* a detailed analysis of the simulated data. The possible theoretical implications of our findings will be discussed in conclusions (see Section 4). To study the validity, we examine the simulation's influence on the original participant categorizations, at both a population (global) and participant (local) level (see Section 3.1). Our findings in the Simulation Influence analysis shows that the end solution maintains the key features of the original categorizations and maintained the original integrity of the data. That is, the evolutionary dynamics does not impose an external structure on the original WCS naming systems. As such, we then create methods to analyze the data to gain insight into the categorizations' evolutionary processes.

5.3.1 Simulation Influence

Global Measure of Change

Over the course of the simulation, the global agreement of a population's categorization A_r is expected to increase as the agents update and learn through their interactions (see Figure 2). ΔA measures the amount of change from the original naming strategies of WCS participants to the converged solution based on A_r:

$$\Delta A = \frac{A_{\bar{r}} - A_0}{A_0} * 100 \tag{5.6}$$

Higher values of ΔA indicate that a greater proportion of chips changed names during the simulation. That is, populations with weak color naming systems (i.e. low A_0) required larger amounts of learning and chip reassignment to reach a shared solution. Conversely, lower values of ΔA imply that the population was initially in high agreement with its color naming convention (i.e. high A_0), and thus did not need much alteration.

The median value for ΔA across all WCS languages is 40%, the mean is 44%, and the standard deviation is 23%. The range of percent change extends from 8% to 123% and the levels of change for 98.14% of all languages were within one standard deviation from the mean. For all WCS language groups, $\Delta A > 0$, which indicates that the simulations were successful in improving the populations' naming system. Improvement is interpreted as the simulations removing previous ambiguity that existed in these populations' categorizations by allowing agents to learn and then develop a shared, common naming system.

The variability in the level of ΔA across the various language groups of the WCS is unsurprising. There was no pragmatic reason for WCS participants to have high expertise across the entire color space or to place equal importance on all chips. Therefore, inter-group variation is to be expected given the improbability that most individuals would have color names for all 320 chips. Hence, the presence of this ambiguity when assigning names in the naming task necessitates that, at the minimum, some degree of learning must take place. Given the differing levels of initial population agreement, A_0, across all WCS language groups, a variable degree of learning to achieve a stable solution took place.

While ΔA is useful in quantifying the total amount of change that a language undergoes during the simulation, it provides little information about how these changes impact each agent's underlying categorization. ΔA reveals what proportion of stimuli were reassigned to different names during the evolutionary process but provides little intuition about whether these changes are a significant departure from each agent's initial categorization. Therefore, to determine whether the simulations preserve the integrity of the original WCS data, we analyze the change to each participant's naming strategy.

Local Measure of Change

In the *mapping task* of the WCS, participants identified focal chip(s) for a set of categories that had previously been elicited by the survey facilitators [68]. A *focal chip*, or *best exemplar*, is a chip (or set of chips) that an individual participant identifies as a best example for a given color word. These chips are those that participants find as the most salient examples of a particular color category. The WCS database includes the complete set of focal chips that were identified by every participant in every language group surveyed [22].

We investigate the focal chips of each participant as a means to understand the influence of the dynamics on the individuals' naming systems. We hypothesis that chips identified as salient to the observers are least likely to change categories. By tracking the change to participant focal chips we analyze the influence of the simulation—low change indicates minimal influence.

A series of steps were completed to isolate the specific data needed for our analysis: first, focal chips that were located on the achromatic axis of the color grid were excluded since the achromatic axis was not included in our simulations; second, only focal chips that were initially identified as focal for a basic color term were considered for analysis; lastly, chips that were identified as focal for a term x but were assigned a different name y in the *naming task* were excluded from analysis due to the contradictory nature of these responses.Because of the way BCTs were defined and WCS data was collected, the name assigned to a focal color chip need not be the name of the BCT for that chip. For example, if a society were to assign *crimson* as the name for a chip in the free listing task but *red* emerged from the analysis as the BCT for that chip, and it was determined that *crimson* was a subcategory of *red*, then *crimson* couldn't be a BCT, because, by definition, a BCT cannot be a subcategory of a larger color category. A more general consideration is that given a category C and finding the chip a that is a focal of that category is a different task than given a and asking the participant to assign the best category name for a from a set of categories including C, even if the set only includes BCTs. This kind of asymmetry is discussed in detail in Jameson & Alvarado (2010). In our study it is the likely reason for why across languages a few focals were observed to switch names from free listing to BCT categorization [61].

Using this focal chip data, we calculate each agent's proportion of focals that retained their original name. The average proportion across the population gives the measure of *focal persistence*, formally defined as follows:

$$P_\ell = \frac{\sum_{i=1}^{N} \frac{|S_i|}{|F_i|}}{N} \tag{5.7}$$

where ℓ is the language number, N is the number of participants in language ℓ, F_i is the set of focal terms for agent i, and $S_i \subseteq F_i$ is the set of focal terms for i which have the same name at the beginning and end of the simulation. The range of P_ℓ is [0,1], where values close to 1 indicate high level of persistence within a language group.

Based on our analysis, the the median value for *focal persistence* across all of the WCS languages is 0.97, the mean value is 0.94, and the standard deviation is 0.11. This result indicates that, on average, 94% of all focal chips identified for a language's basic color terms retained their original category name throughout the simulation—i.e. important chips retain original names and learning takes place primarily on non-focal chips (see Figure 6).

Hence, despite the changes in each language's categorization during the simulation (evident from the percent change measures in Section 3.1.1), the change to the underlying structure of these naming systems is minimal. In other words, there is support that the communi-

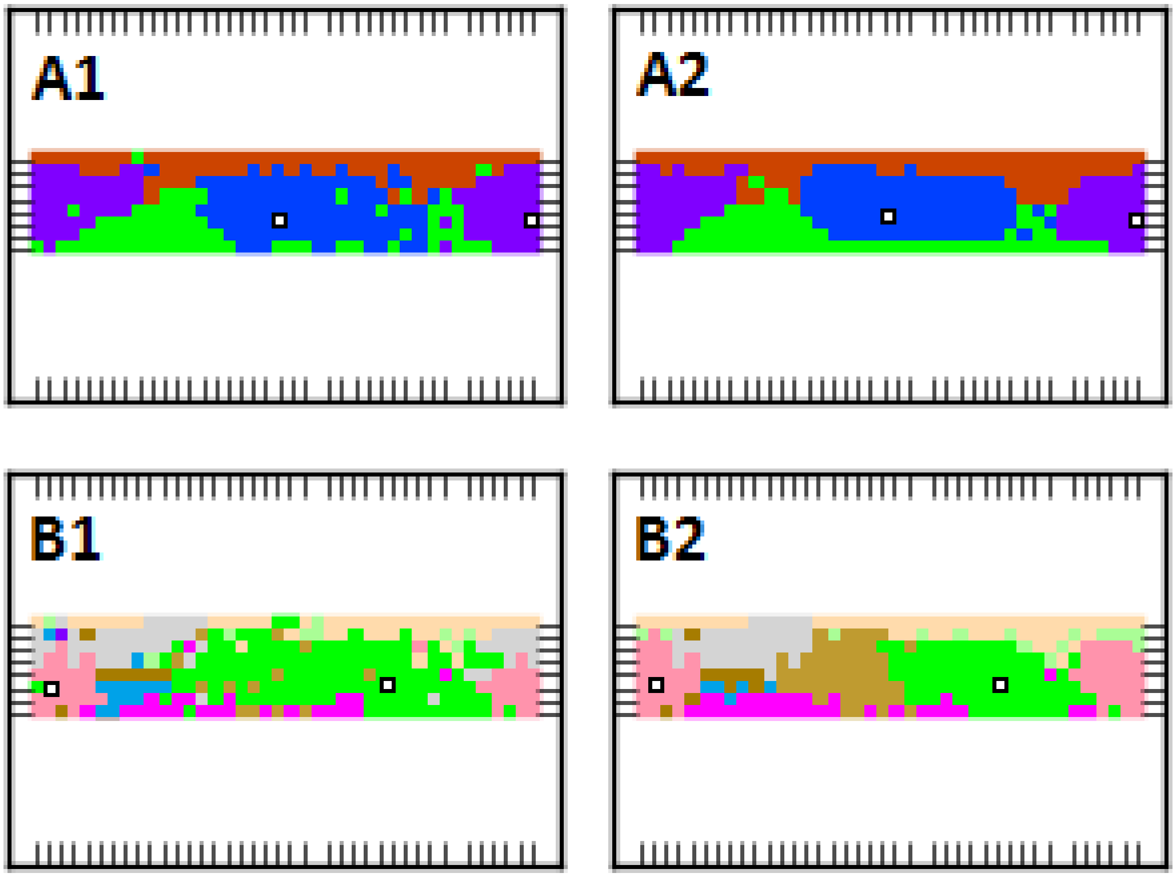


Figure 5.8: Naming systems for agents in Languages 74 (A*) and 95 (B*) at the beginning of the simulation (*1) and at the end of the simulation (*2) with markers for focal chips identified by those agents in the mapping task of the WCS. Focals are indicated using a white square with a black outline. Language 74 (A) is an example of a language with a small $\%\Delta A$. Language 95 (B) had large $\%\Delta A$.

cation mechanism used in these simulations is not imposing an external structure on the original WCS data but is instead improving the categorization across the populations while maintaining important features of agents' categorizations.

5.3.2 Simulation Analysis

To study the evolution of categorization, it is important to understand both the features of categories and how these features play into the evolution of categories. In this analysis, categories are defined as a collection of stimuli that share the same name. While chips located at a boundary of a category and at the center share the same name, there are important differences between the two (e.g. chips on the boundary are subject to persistent name changes while central chips are likely not). However, with no standard metric for identifying category boundaries, we develop a new probabilistic measure to identify chip location within a category. Knowing how each location of a category evolves helps us understand how a category is changing as a whole. Performing this location-based analysis will provide insight as to what evolutionary processes are driving these observed category changes.

Identifying Boundary Chips

The WCS provides individualized data, per participant. There is no principled way to aggregate these categorizations into one unified, population-wide naming strategy. Therefore, the concept of *boundary* is difficult to define in a deterministic way. Hence, we develop a probabilistic approach to identifying category boundaries.

For any chip j in the 320-chip stimulus set, we first calculate the *boundary value* of chip j, which is the proportion of chips within 1 *k-sim* that have the same name as chip j.

$$B_{\ell,i}(c) = \frac{|S_c|}{|K_c|} \tag{5.8}$$

where ℓ is the language number, i is the numeric ID of the agent, c is the chip number, K_c

is the set of chips within 1 *k-sim* of chip c, and $S_c \subseteq K_c$ is the set of chips within 1 *k-sim* of c that have the same name as c according to agent i. *Boundary value* increases as a chip becomes more central to a category—the higher the measure, the less likely that chip is a boundary chip.

Let the average *boundary value* for a given chip c across all participants in language ℓ be

$$\bar{B}_\ell(c) = \frac{\sum_{i=1}^{N} B_{\ell,i}(c)}{N}. \tag{5.9}$$

We then employ the following method to identify the likelihood that chip j is on the "boundary" of a color category. We define $f_\ell : [0,1] \to [0,1]$ to be the probability density function of *boundary values* for language ℓ such that $f_\ell(b) = P(\bar{B}_\ell = b)$, for $b \in [0,1]$. This function is constructed using aggregated frequency data from the *boundary values* of all 320 chips in the stimulus set.

Let $F_\ell : [0,1] \to [0,1]$ be the function that converts *boundary values* to *boundary probabilities* defined by

$$F_\ell(b) = \int_b^1 f_\ell(b)db \tag{5.10}$$

From this measure, we can then define the probability that a chip is part of a category "boundary" by

$$BP_\ell(c) = F_\ell(\bar{B}_\ell(c)) \tag{5.11}$$

Through the *border probability* measure, $BP_\ell(c)$, we can obtain an estimate for the geographic location within a color category. The higher the *border probability*, the more likely that a chip is on a category boundary. The lower the *border probability*, the more likely that a chip is central to a category.

Categorization Analysis

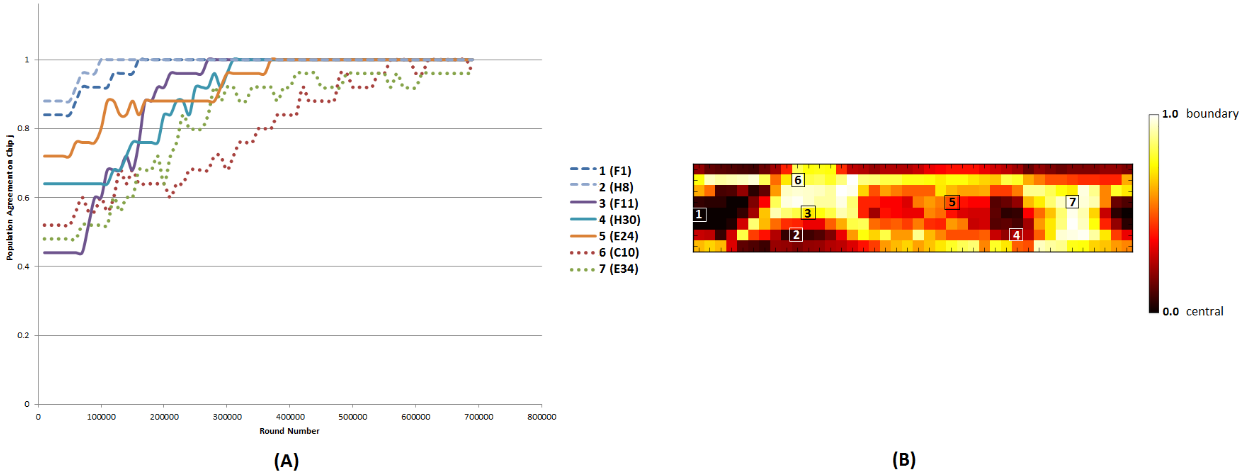


Figure 5.9: Simulations for WCS Language 23: (A) Plot of chip agreements ($P_{i,\alpha}(c)$) seven chips. (B) Heat map depicting the boundary probability ($BP_\ell(c)$) of each chip, organized on the 320 chip color grid. Boxed numbers on heat map correspond to the chips identified in the plot. From the results depicted in (B), we classify chips 1 and 2 as having *high* border probabilities, chips 3, 4, and 5 as having *moderate* border probabilities, and chips 6 and 7 as having *low* border probabilities. From (A), we observe that *low* border probability chips are the first to reach total agreement ($P_{i,\alpha}(c) = 1$), then the *moderate* border probability chips, and lastly the *high* border probability chips. This suggests that at a social level border chips are the last to be learned.

We hypothesize that some chips will evolve to a stable naming convention faster than others. Namely, central category chips, and possibly focal chips, due to their high salience, will reach full agreement more quickly than category boundary boundaries. Firstly, through the measure $BP_\ell(c)$, we can calculate the probability that any given chip is on a category boundary. Secondly, we then track the chip agreement by calculating the proportion of the population that uses the most common name for each chip, remeasured at a fixed interval for the duration of the simulation. Chips reach full agreement when the proportion equals 1. Together these two measures allow for a location-based analysis of categorization evolution.

To estimate *convergence rate* of a chip (i.e. the time for a population to reach full agreement on the chip name) we calculate mean of proportions,

$$CR_\ell(c) = \frac{\sum_{i=1}^{\bar{r}} P_{i,\alpha}(c)}{\bar{r}} \tag{5.12}$$

where $\bar{r}$ is the total number of game rounds, i is a round number, c is a chip number, α is the most frequently used name for chip c, and $P_{i,\alpha}(c)$ is proportion of agents in a population who assign name α to chip c. $P_{i,\alpha}(c)$ equals 1 when the whole population agrees on what name to call c. Therefore, the higher value of $CR_\ell(c)$, the sooner in the simulation that chip achieves total agreement.

To study the relationship between chip location within a category and the rate of convergence to full agreement, we calculate $corr(BP_\ell, CR_\ell)$ across all 320 color chips in the stimulus set. This provides an aggregated measure of how boundary probabilities relate to convergence time for each language.

The correlations between BP_ℓ and CR_ℓ for all language groups were negative, indicating that chips that are likely to be on a category boundary are also likely to converge slower—i.e. reach total agreement later in the simulation. Figure 7 presents a specific example of this finding.

Additionally, 88 percent of all languages have a correlation strength of 0.5 and above with close to 50 percent of language groups at a strength of 0.7 and above (see Figure 8). Therefore, this measure of the relationship between border probability and convergence rate is non-trivial for a majority of the languages tested in this paper.

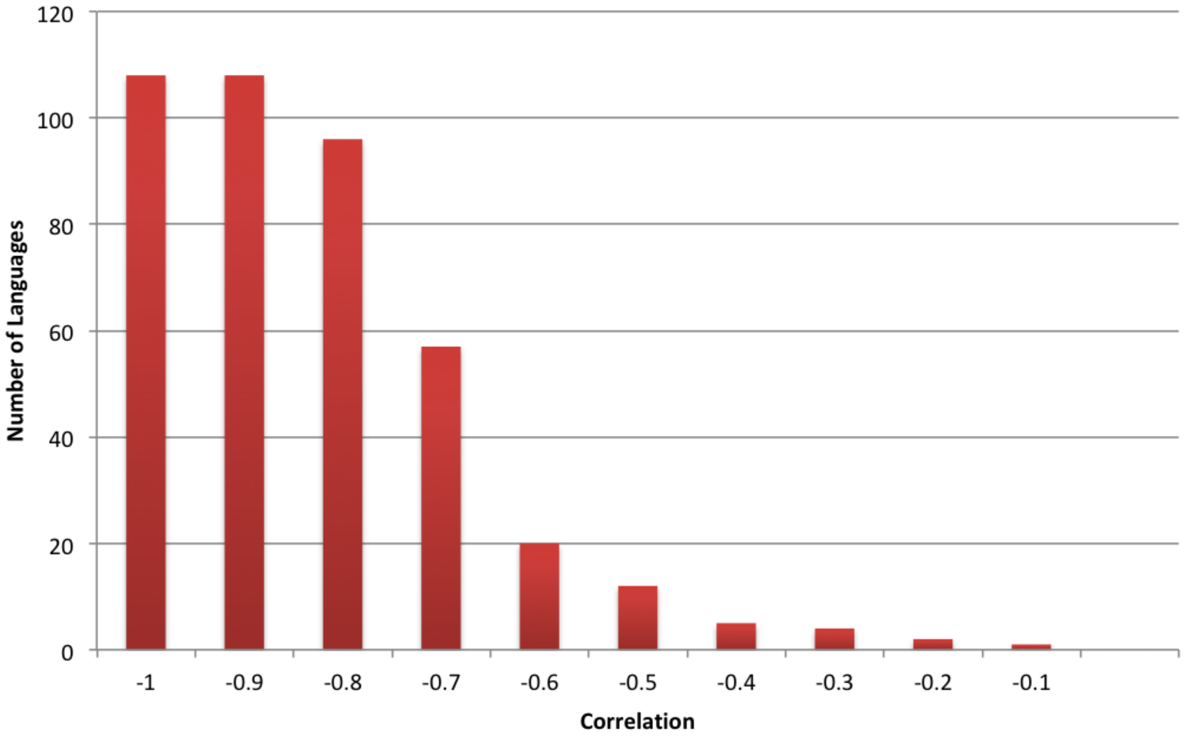


Figure 5.10: Cumulative histogram of number of languages that have a boundary-convergence correlation measure less than the correlations along the x-axis.

5.4 Conclusion

The goal of taking a simulation-based approach to color category evolution, beyond developing and refining a “workable” model, is to be able to test and evaluate theories of color categorization. Two of the most prominent competing theories regarding color category evolution are the *Emergence* and *Partition Hypotheses*. The measures detailed in *Results* (see Section 3), provides some insight into an idealized approximation of categorization evolution which can lend its support to one of the theories.

By analyzing the structure of the data in the end solution, we observed that almost all focals maintained their original categorizations, despite large amounts of system change. This may seem contradictory for languages with low initial agreement. Why would foci persist when there seems to be an overall weak notion of color within the population? In communities where color is not a salient concept, it is reasonable to see highly varied and sometimes inconsistent naming strategies, particularly in regions of the color space they had no pragmatic need to name. The *mapping task*, on the other hand, allowed participants to choose chips that best represented color categories, meaning they likely chose chips that were highly salient. These salient chips thus served as “anchors” to the category and during the simulations, as the agents engage in the learning dynamic, they gain expertise in these regions of confusion around the focal chips. Hence, the focal chips for a category persist in the final naming system and learning takes place at the category boundaries.

In the same vein, the analysis of the correlation between border probability measure, $BP_\ell(c)$, and chip convergence rate, $CR_\ell(c)$, reveals a substantial relationship for most languages. Therefore, there is evidence that chips located in the center of a category will reach full agreement more quickly than border chips. In other words, results indicate that the area of salience and expertise grows outwards from the focal chips to the ambiguous chips—a society will gain expertise for central chips first and with continuous communication, the expertise will move outwards towards the category boundary.

Taken together, these results show that central chips have high salience and anchor the categorizations throughout the evolutionary processes of the simulation. These results, therefore, lend support in favor of the *Emergence Hypothesis* of color evolution which states that categories begin with highly salient points and then extend outwards.

5.5 Discussion

This paper was a preliminary exploration of *ColorSims 2.0* as a tool to analyze natural color categorizations. With minimal perceptual and memory assumptions for agents, we were able to evolve color naming systems, using both randomized and non-randomized agent populations from the WCS, to a stable solution using the two communication games, *Discrimination-similarity game* and *2-player teacher game*. By inducing these systematic interactions, the concept of color was strengthened within a population and evolved into a more salient, stable convention that had high social agreement. Using WCS as a test case was crucial to our ability to verify the validity of our approach. As hoped, the impact of these simulations on the original WCS data was minimal and our model imposed very little influence on the underlying structure of the categorizations. Yet, working with a such a data set carries its own limitations.

WCS data provides a set of pre-existing categorization schemas with names for the entire set of stimuli—following Berlin & Kay's theory that BCTs perfectly partitioned the color space. With each society's fully labeled stimuli sets, Berlin & Kay chose a "best BCT" for each chip by taking the chip's modal color term. There, however, may exist little agreement about such a method, calling into question the usefulness of such "bests" in actual communicative practice. Similar concerns arise for the selection of a focal chip for a BCT.[3] Though our analysis takes such possible, implicit limitation into consideration, there is not much that can be done with the data that is available through the WCS. While these considerations may limit the extent to which we can weigh in on theories of the evolution of color categorization, our results using WCS data is both promising and gives hope for further exploration.

Theories of color categorization make claims about how categories evolve from their conception. While using WCS data limited our understanding about how these categories evolved to that point, in later iterations we will update the evolutionary dynamics in such a way that better captures the evolutionary processes described by these theories. For example, initializing agent categorizations with a few salient points and allowing the other space to be learned. Future research will include developing a measure to examine the strength of a color categorization to provide a more detailed understanding of how categorizations are changing at the population level over time—semantic drift. Additionally, this measure will provide the basis for a new criterion used to group language communities in cross-population analysis. There have been methods developed to categorize naming systems according to stages of evolution, but these methods lack information about how established a language is within its evolutionary stage. Performing this analysis using simulation-generated data would afford us the means and opportunity to draw specific conclusions about where a language is in its evolutionary process.

5.6 Future studies

The trends present in the MacLaury data (MCS) can give interesting insight into the evolution of linguistic conventions by studying populations that hail from the same family language but separated into different groups. This contrasts with the data available from the WCS that consists of independent language groups and the evolutionary patterns were approximated by imitating communication dynamics [41]. To assess the evolutionary patterns present in the MCS, this study utilizes the two-player teacher game [61] with population dynamics: Birth death, Replicator dynamics, Best response.

An earlier study analyzed the evolution of naming conventions on a changing population [**?**] using the birth-death dynamics on a random population of agents. This study will initialize agents with the WCS and MCS data using the Colorsims 2.0 framework with *(i)* a higher dimensional color space that more realistically approximates human color perception, *(ii)* real observer population data taken from the WCS, *(iii)* a measure to assess the stability of a color naming convention [39].

5.6.1 Methods

Taking a subset of the Mesoamerica groups (see figure 5.11) that hail from the same language family, we look at the

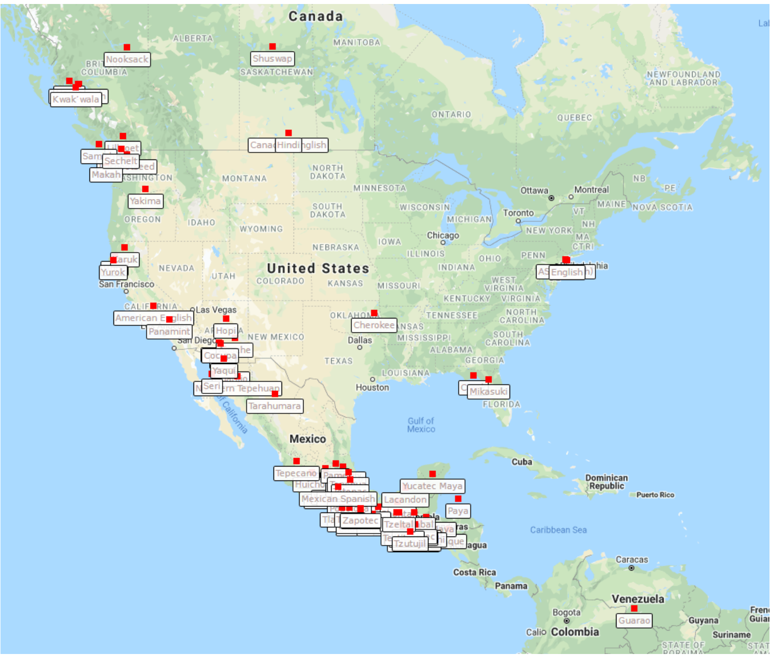


Figure 5.11: The Mesoamerican groups studied.

Initially, all MCS languages we have available will be taken to a stable convention and internal markers of stability will be measured based on the methods of this paper. This will ensure that the naming systems maintained their key features under the two-player teacher game's evolutionary process - i.e. no external structure was imposed on the name system. A thorough analysis will compare results from this paper to the results from MCS data before testing the evolutionary patterns between the Mesoamerican tribes using a variety of population dynamics. The two-player teacher game will be implemented with a gradually changing population of agents initialized with the MCS data to see which most closely resembles the data from the data.

Population dynamics

Birth Death At a given rate of n a new members with a random naming system will be initialized and at rate m, a random agent will be eliminated as outlined in Park et. al. [101].

Best Response Initially, a modal map for the population is generated: a matrix of 320 arrays with each array representing a list of names each participant chose for the corresponding chip. When the game begins, agents will consult the best strategy in the group for assigning names to the chosen color chips based on the population modal map. Their own naming strategy will be only one input in the population's modal map. Upon updating their naming strategies, each agent has the potential to influence the population's modal map if their update changes the mode for the corresponding color chip.

Replicator dynamics Initially, a rank order of agents based on level of expertise will be generated (i.e. the closest agent to the model map is the most expert in the group). In each period, each agent is randomly matched with another agent and updating happens as described in [72]. An agent's fitness f is determined by level of expertise using strategy A or B (in this case, based on their naming schema, the agent who is more expert will have the better strategy). The rate of replication will be calculated every round using the replicator equation:

$$\frac{dx_A}{dt} = x_A * (f_A(x_A, x_B) – \bar{f}(x_A, x_B)) \tag{5.13}$$

Where A is $\frac{dx_A}{dt}$ is the growth rate of strategy A, x_A is the current frequency (proportion) of strategy A in the population (indicating how many A -players can reproduce), $f_A(x_A, x_B)$ is the payoff fitness of an agent using A strategy, $\bar{f}(x_A, x_B)$ is the average fitness of the population, and, $(f_A(x_A, x_B) – \bar{f}(x_A, x_B))$ is an A -player's fitness relative to the average fitness (i.e. the key property: More successful strategies grow faster) if there is a non-zero population of agents playing strategy A ($x_A > 0$) and the fitness of an agent playing A is above average ($f_A > \bar{f}$), then the population that plays strategy A will increase ($\frac{dx_A}{dt} > 0$).

Chapter 6

Concluding thoughts

With advances in Machine Learning (ML) and great improvements in computational power available to researchers, studying social conventions scientifically is possible more than ever before. Many researchers, including Joe et. al. [62] have employed ML techniques to uncover features in Linguistic conventions, particularly color naming, with much success. Due to the availability of the World Color Survey (WCS) data, color naming enjoyed attention from various department, all with similar goals: to uncover important features of language. Insights gleaned from these effort build upon each other, creating an active area of research. This is yet to be the case for a similar phenomenon, ideology. ML has provided researchers with the ability to study ideology scientifically, making minimal theoretical assumptions–particularly those tied to departmental biases and goals.

Ideology is a complex topic of study. The same is true for language. Both deal with how meaning comes about. Language is the assignment of meaning to sounds–that then become words– while ideology is assignment of meaning to actions, beliefs, and words. Studying how linguistic meaning comes about is a complicated and difficult undertaking, so is studying how ideologies form. Less difficult however, is comparing color naming or the lexicon between two language groups. The same is true of comparing the concept of sins between religions groups. Both the concept of color and sin are universal across groups and by studying these universal examples, communities of researchers can amass much insights into these phenomena, with minimal bias. This dissertation outlines ways of doing so.

In Chapter 2, Kirbi Joe and I employ sophisticated ML techniques, namely infinite non Bayesian mixture models to make variational inference that uncovered 18 universal patterns in the color naming of the WCS participants. To do so, we took key features of the color naming task to do so: influence of culture, participants biological ability to discern between colors, and the features of the color space. Here, rather than bringing in our researcher biases, we took these fundamental aspect of this phenomenon as key assumptions of the model. Explicitly, the assumption of the color naming task is that it is a social activity, bounded by each linguistic group's culture where the millions of colors that each participant can observe is assigned to a small finite set of words (e.g. dark, light, red) that are communicated on. Yet, the names are not merely given at random, the closer two colors are in the stimulus space, the higher the probability of them assigned the same name. These assumptions alone allowed us to transform each participant's data to a binary vector of neighborhood judgments. By doing so, the ML algorithm was able to compare participant judgments and assign them to clusters based on level of similarity. We developed a novel metric for measuring two individual's schematic similarities, independent of their language. This approach became a road map to the study of another, similar social convention: ideology.

In Chapter 3, we extend the methodology outlined in the second chapter to study ideologies of various political, religious, and cultural groups. Instead of starting from a definition of ideology, we make the assumption that these self identified group possess an ideology. Our task is then to compare the beliefs between these groups. This approach allows us to transform a complicated task to a far more simple one. The key challenge in studying ideology is that there is no standard data set like the WCS. So, over the course of two years, a set of cultural questions or “items” were created to allow for gathering data. With the sample data gathered and using computational and mathematical techniques such as Cultural Consensus Theory and the Dirichlet process, we were able to discover 4 main ideological groups. These groups had unique features: one group was marked by their moderate and non supernatural beliefs as well as low regard to sunshine, another moderate group had strong beliefs in an assortment of supernatural ideas and were big sports fans, the group of liberals had a high level of anti-conservative views, while the conservatives possessed traditional religious beliefs and were driven by ideas about their country. This data set of binary (i.e. True, False) responses, allowed for the clustering of a diverse participant population. But a non-binary data set could allow for an even richer exploration of differences.

In chapter 4, we zoom in even further and look at religious groups. Arguably, religious ideology is the easiest to identify type ideology. All large religions today have a lengthy oral and written history, congregate on a regular basis, and often have visible and unique features (e.g. dress, symbols, diet, rituals). We study religious ideology by examining a universal concept central to all religions: sins, particularly the very famous Seven Deadly sins. Using chess scaling algorithms we investigated the each group’s ranking of the sins from worst to least worst. This chapter consists of two studies. First study is the testing of the Elo system to aggregate group rank data. We gathered data on the rank ordering of the seven sins and then of an additional five sins–with the second set having three of the previous sins in common from the first set to allow for the ordering of the twelve total sins. The Elo model successfully aggregated group rankings while our clustering algorithm found four major groups. The second study then examines the extent to which Christian beliefs are known. Here, we ask Atheists and Christians to each respond as if they were devout Christians. The findings showed that Atheists were unable to do so while the Christian successfully completed the task.

Lastly, Chapter 5 examines evolutionary game theoretic models to study the underlying dynamics of the formation of convention–particularly, linguistic color naming. Reinforcement learning (RL) as an evolutionary model is an appropriate one to model social conventions. It mimics the cultural processes of learning (i.e. socialization) and specifically, the implementation of RL in the Two Player Teacher game introduced by Komorova el. al [72]. Here, we were able to model real observers and show that the pragmatics of language alone, that is, the need to communicate, results in the development of linguistic conventions. In this case, color categories emerged from noise. Having successfully shown the model to work on random data, we initialized the system with WCS participant data and evolved each group's color naming schema to a stable convention. We developed our own notion of stability and were able to see how the system influenced each participant's and group's data. We found that the Two Player Teacher game preserved the most salient color items' assignments and made imposed little instrumental bias.

The insights from Chapter 3 and Chapter 4 as well as the success of the evolutionary approach in Chapter 5 can inspire future work in the study and modeling of ideology. The following section consists of some early thoughts on this effort.

6.1 Possible Evolutionary models

From the studies conducted in this dissertation, we see that groups can have different levels of shared ideas, beliefs, norms. Importantly, these social conventions can play an important role in the group's survival: enforces cooperation, identifies right and wrong, provide rules for decision making on important topics such as marriage, inheritance, rulership, division of labor and resources, etc. In the case of more lasting ideologies, these norms can be embedded in supernatural forces beyond that of human control: the divine, natural laws of nature, ideals such as notions of justice or fairness. When such social conventions develop into "the" way to pursue life, they can take on the form of an ideology and provide a worldview that

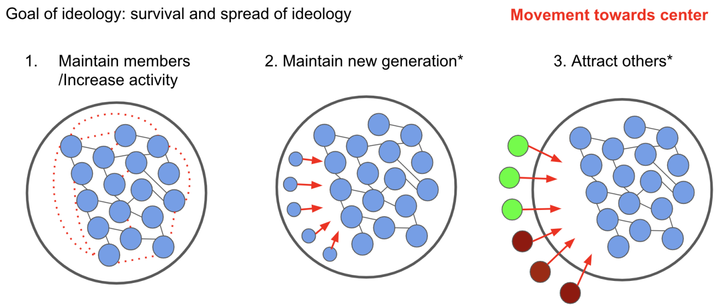

Goal of ideology: survival and spread of ideology
Movement towards center
1. Maintain members /Increase activity
2. Maintain new generation*
3. Attract others*

Figure 6.1: Possible strategies of ideologies. An ideology can use all options but the original group it was developed may inhibit the use of some of the options (like Hindus who rely on birth).

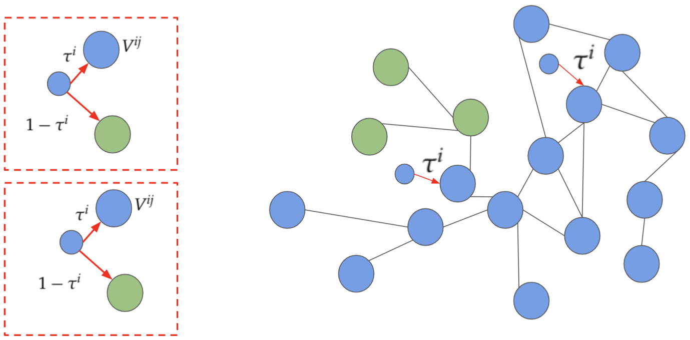


Figure 6.2: The probability can be a function of the centrality of parent.

minimizes inquiry and present its conventions as indisputable, universal truths.

Ideology functions through its members. While an ideology depends on follow-ship, after a critical mass of followers is reached and its laws are stabilized to a sufficient extent, an ideology may not be directly influenced by the number of members. With that said, an ideology for a small society, nomad society, or large metropolis may differ in its rules, norms, and expansion mechanism. In general, there are a few option that an ideology can use to strive (figure 4): 1. Increase in follow-ship by birth 2. Increase follow-ship by conversion from the outside 3. Maintaining members (i.e. make it difficult to leave group)

Some consideration can be the likelihood of cultural transmission for the new generation (figure 5).

Each ideology can be in competition with the other (figure 6).

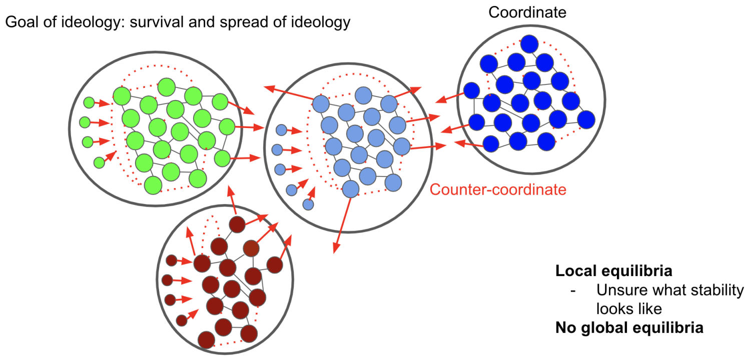

Goal of ideology: survival and spread of ideology
Coordinate
Counter-coordinate
Local equilibria
- Unsure what stability looks like
No global equilibria

Figure 6.3: Competition between ideologies.

Ideologies arise through interactions of members yet, after a critical mass of members is reached, it is not directly caused by it. AS it is with any group, as norms evolve, consensus about them changes and the ideology must also adapt and evolve to maintain its members to survive. If the new set of norms are distant enough from the original message, a cleavage may form in the group resulting in a new ideology. What seems to matter greatly are the rules that govern its social conventions. If these rules are rigid, groups may not attract many members or be able to maintain existing ones yet. Lax rules on the other hand can limit the formation of a unique group identity–i.e. anything goes. Ideological groups can use a number of different strategies: Inter relatedness of tenants, Empirical relevance, Utility of ideology,, Adaptability (DA), Tolerance of "other" , Degree of commitment required, etc.

Insights from Chapter 3 do also point to three other possible mechanisms. First, some groups form in opposition to another dominant group as was the case with the liberal cluster. Second, a group can form because of their lack of belief. Lastly, a group can form around supernatural ideas that do not belong to a single religious group but rather an eclectic set of beliefs.

Bibliography

[1] K. Agrawal and W. H. Batchelder. Cultural consensus theory: Aggregating signed graphs under a balance constraint. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 53–60. Springer, 2012.

[2] L. Althusser. Ideology and ideological state apparatuses. *Lenin and philosophy and other essays.*, 1971.

[3] K. Ameriks and D. M. Clarke. *Aristotle: Nicomachean Ethics*. Cambridge University Press, 2000.

[4] D. M. Amodio, J. T. Jost, S. L. Master, and C. M. Yee. Neurocognitive correlates of liberalism and conservatism. *Nature neuroscience*, 10(10):1246–1247, 2007.

[5] R. Anders and W. H. Batchelder. Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, 56(6):452–469, 2012.

[6] R. Anders and W. H. Batchelder. Cultural consensus theory for the ordinal data case. *Psychometrika*, 80(1):151–181, 2015.

[7] R. Anders, Z. Oravecz, and W. H. Batchelder. Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, 61:1–13, 2014.

[8] T. Aquinas. Summa theologiae, 2005.

[9] W. Bagehot. Physics and politics - or thoughts on the application of the principles of natural selection and inheritance to political society. 1873.

[10] J. Barnes. *Complete works of Aristotle, volume 1: The revised Oxford translation*, volume 1. Princeton University Press, 1984.

[11] A. Baronchelli, T. Gong, A. Puglisi, and V. Loreto. Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences USA*, 107(6):2403–2407, 2010.

[12] W. H. Batchelder, R. Anders, and Z. Oravecz. Cultural consensus theory. pages 1–64, 2018.

[13] T. Belpaeme and J. Bleys. Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior*, 13(4):293–310, 2005.

[14] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[15] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[16] B. Berlin. *Basic Color Terms: Their Universality and Evolution.* University of California Press, 1969.

[17] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[18] J. Bleys and T. Belpaeme. Explaining universal colour categories through a constrained acquisition process. In *BNAIC*, pages 317–318. Citeseer, 2005.

[19] D. Centola and A. Baronchelli. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences USA*, 112(7):1989–1994, 2015.

[20] J. Chiao, V. Mathur, T. Harada, and T. Lipke. Neural basis of preference for human social hierarchy versus egalitarianism. *Annals of the New York Academy of Sciences*, 1167(1):174–181, 2009.

[21] A. Clericuzio. Elements, principles, and corpuscles. 2000.

[22] R. S. Cook, P. Kay, and T. Regier. The world color survey database. In *Handbook of categorization in cognitive science*, pages 223–241. Elsevier, 2005.

[23] B. De Vylder and K. Tuyls. How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of theoretical biology*, 242(4):818–831, 2006.

[24] S. D. Driver. *John Cassian and the reading of Egyptian monastic culture*. Routledge, 2013.

[25] E. Durkheim. The rules of sociological method. 1937.

[26] K. Emmet. 'ideology' from destutt de tracy to marx. *Journal of the History of Ideas*, 40(3):353–368, 1979.

[27] N. Fairclough. Language and power. 2001.

[28] N. Fider and N. L. Komarova. Quantitative study of color category boundaries. *JOSA A*, 35(4):B165–B183, 2018.

[29] N. Fider, L. Narens, K. A. Jameson, and N. L. Komarova. Quantitative approach for defining basic color terms and color category best exemplars. *JOSA A*, 34(8):1285–1300, 2017.

[30] N. A. Fider and N. L. Komarova. Differences in color categorization manifested by males and females: a quantitative world color survey study. *Palgrave Communications*, 5(1):1–10, 2019.

[31] S. D. Findlay and P. Thagard. Emotional change in international negotiation: Analyzing the camp david accords using cognitive-affective maps. *Group Decision and Negotiation*, 23(6):1281–1300, 2014.

[32] A. Finlayson. Rhetoric and the political theory of ideologies. political studies. 60(751-767), 2012.

[33] M. Freeden. Ideology: A very short introduction. *Lenin and philosophy and other essays.*, 2003.

[34] M. Freeden. Social and psychological bases of ideology and system justification. *Political Psychology*, 31:479–482, 2010.

[35] L. Geiger. *Contributions to the History of the Development of the Human Race: Lectures and dissertations*, volume 12. Trübner & Company, 1880.

[36] S. J. Gershman and D. M. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[37] W. E. Gladstone. *Studies on Homer and the Homeric age*, volume 1. Oxford University Press Oxford, 1858.

[38] J. Glanvill. The vanity of dogmatizing: or confidence in opinions. manifested in a discourse of the shortness and uncertainty of our knowledge, and its causes; with some reflexions on peripateticism; and an apology for philosophy. pages [32], 250, [6] p. ;, 1661.

[39] M. Gooyabadi and K. Joe. Colorsims 2.0: an extension to the python package for evolving linguistic color naming conventions applied to a population of agents. Technical report, IMBS Technical report series, MBS 18-02, Institute for mathematical ..., 2018.

[40] M. Gooyabadi, K. Joe, and L. Narens. Further evolution of natural categorization systems: an approach to evolving color concepts. *JOSA A*, 36(2):159–172, 2019.

[41] M. Gooyabadi, K. Joe, and L. Narens. Further evolution of natural categorization systems: an approach to evolving color concepts. *J. Opt. Soc. Am. A*, 36(2):159–172, Feb 2019.

[42] J. Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

[43] P. L. Hammack. Narrative and the cultural psychology of identity. *Personality and social psychology review*, 12(3):222–247, 2008.

[44] J. Hampton. *Hobbes and the social contract tradition*. Cambridge University Press, 1988.

[45] C. L. Hardin and L. Maffi. *Color categories in thought and language*. Cambridge University Press, 1997.

[46] P. K. Hatemi, N. A. Gillespie, L. J. Eaves, B. S. Maher, B. T. Webb, A. C. Heath, S. E. Medland, D. C. Smyth, H. N. Beeby, S. D. Gordon, et al. A genome-wide analysis of liberal and conservative political attitudes. *The Journal of Politics*, 73(1):271–285, 2011.

[47] E. R. Heider. Universals in color naming and memory. *Journal of experimental psychology*, 93(1):10, 1972.

[48] E. R. Heider and D. C. Olivier. The structure of the color space in naming and memory for two languages. *Cognitive psychology*, 3(2):337–354, 1972.

[49] S. Heshmati, Z. Oravecz, S. Pressman, W. H. Batchelder, C. Muth, and J. Vandekerckhove. What does it mean to feel loved: Cultural consensus and individual differences in felt love. *Journal of Social and Personal Relationships*, 36(1):214–243, 2019.

[50] D. D. Hoffman and M. Singh. Computational evolutionary perception. *Perception*, 41(9):1073–1091, 2012.

[51] T. Homer-Dixon, J. L. Maynard, M. Mildenberger, M. Milkoreit, S. J. Mock, S. Quilley, T. Schröder, and P. Thagard. A complex systems approach to the study of ideology: Cognitive-affective structures and the dynamics of belief systems. *Journal of Social and Political Psychology*, 1(1), 2013.

[52] F. N. House. Pareto in the development of modern sociology. *J. Soc. Phil.*, 1:78, 1935.

[53] M. C. Hughes and E. B. Sudderth. Memoized online variational inference for dirichlet process mixture models. Technical report, BROWN UNIV PROVIDENCE RI DEPT OF COMPUTER SCIENCE, 2014.

[54] D. Hume. *A treatise of human nature*. Courier Corporation, 2003.

[55] D. Hume et al. *An enquiry concerning human understanding: A critical edition*, volume 3. Oxford University Press, 2000.

[56] W. James. Principles of psychology - the theories of instincts. 1890.

[57] K. Jameson and R. G. D'Andrade. 14 it's not really red, green, yellow, blue: an inquiry into perceptual color space. *Color categories in thought and language*, page 295, 1997.

[58] K. Jameson, N. Komarova, S. Tauber, and L. Narens. New results on simulated color categorization behaviors using realistic perceptual models, heterogeneous observers and pragmatic communication constraints., 2011. In *Presentation at The 17th annual meeting of the Cognitive Science Association for Interdisciplinary Learning*, 2011.

[59] K. A. Jameson and N. L. Komarova. Evolutionary models of color categorization. i. population categorization systems based on normal and dichromat observers. *JOSA A*, 26(6):1414–1423, 2009.

[60] K. A. Jameson and N. L. Komarova. Evolutionary models of color categorization. ii. realistic observer models and population heterogeneity. *JOSA A*, 26(6):1424–1436, 2009.

[61] K. A. Jameson and N. L. Komarova. Evolutionary models of color categorization based on realistic observer models and population heterogeneity. *Journal of Vision*, 10(15):26–26, 2010.

[62] K. Joe and M. Gooyabadi. A bayesian nonparametric mixture model for studying universal patterns in color naming. *Applied Mathematics and Computation*, 395:125868, 2021.

[63] T. Jones. Pope and translations of plutarch's moralia. *Translation and Literature*, 12(2):263–273, 2003.

[64] J. T. Jost. "elective affinities": On the psychological bases of left–right differences. *Psychological Inquiry*, 20(2-3):129–141, 2009.

[65] R. Kanai, T. Feilden, C. Firth, and G. Rees. Political orientations are correlated with brain structure in young adults. *Current biology*, 21(8):677–680, 2011.

[66] L. Kaufman and P. J. Rousseeuw. Finding groups in data: An introduction to cluster analysis–john wiley & sons. *Inc., New York*, 1990.

[67] P. Kay, B. Berlin, L. Maffi, W. Merrifield, et al. Color naming across languages. *Color categories in thought and language*, 21(2), 1997.

[68] P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook. *The World Color Survey*. CSLI Publications Stanford, 2009.

[69] P. Kay and L. Maffi. Color appearance and the emergence and evolution of basic color lexicons. *American anthropologist*, 101(4):743–760, 1999.

[70] P. Kay and C. K. McDaniel. The linguistic significance of the meanings of basic color terms. *Language*, pages 610–646, 1978.

[71] N. L. Komarova and K. A. Jameson. Population heterogeneity and color stimulus heterogeneity in agent-based color categorization. *Journal of theoretical Biology*, 253(4):680–700, 2008.

[72] N. L. Komarova, K. A. Jameson, and L. Narens. Evolutionary models of color categorization based on discrimination. *Journal of Mathematical Psychology*, 51(6):359–382, 2007.

[73] J. Konstantinovsky. *Evagrius Ponticus: the making of a gnostic*. Routledge, 2016.

[74] S. C. Levinson. Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, 10(1):3–55, 2000.

[75] D. Lewis. Languages and language. reprinted in philosophical papers (vol. i), 1975.

[76] D. Lewis. *Convention: A philosophical study.* John Wiley & Sons, 2008.

[77] D. T. Lindsey and A. M. Brown. Universality of color names. *Proceedings of the National Academy of Sciences*, 103(44):16608–16613, 2006.

[78] D. T. Lindsey and A. M. Brown. World color survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences*, 106(47):19785–19790, 2009.

[79] D. T. Lindsey, A. M. Brown, D. H. Brainard, and C. L. Apicella. Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, 25(18):2441–2446, 2015.

[80] J. Locke. *Two treatises of government.* Yale University Press, 2008.

[81] V. Loreto, A. Mukherjee, and F. Tria. On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, 109(18):6819–6824, 2012.

[82] V. Loreto and L. Steels. Social dynamics: Emergence of language. *Nature Physics*, 3(11):758, 2007.

[83] R. D. Luce. Luce's choice axiom. *Scholarpedia*, 3(12):8077, 2008.

[84] J. Luttinen. Bayespy: variational bayesian inference in python. *The Journal of Machine Learning Research*, 17(1):1419–1424, 2016.

[85] K. Mannheim. Ideology and utopia: An introduction to the sociology of knowledge. 1949.

[86] K. Marx. *Capital: volume one.* Courier Dover Publications, 2019.

[87] K. Marx and F. Engels. *The communist manifesto.* Yale University Press, 2012.

[88] J. L. Maynard. A map of the field of ideological analysis. *Journal of Political Ideologies*, 18(3):299–327, 2013.

[89] W. McDougall. An introduction to social psychology. 1908.

[90] W. H. McGovern. From luther to hitler; the history of fascist-nazi political philosophy - ch ix "irrationalism and the irrationalists". *American Political Science Review*, 35(5):400–452, 1941.

[91] W. H. McGovern. From luther to hitler; the history of fascist-nazi political philosophy - ch x "the social darwinists and their allies". *American Political Science Review*, 35(5):400–452, 1941.

[92] D. McLellan. Ideology. 1995.

[93] M. Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003.

[94] S. Mithen. The hunter—gatherer prehistory of human—animal interactions. *Anthrozoös*, 12(4):195–204, 1999.

[95] E. J. Mundy III. *Gerard David Studies.* Princeton University, 1980.

[96] L. Narens, K. A. Jameson, N. L. Komarova, and S. Tauber. Language, categorization, and convention. *Advances in Complex Systems*, 15(03n04):1150022, 2012.

[97] M. Ni, E. B. Sudderth, and M. Hughes. Variational inference for beta-bernoulli dirichlet process mixture models.

[98] D. R. H. A. J. Norval, Y. Stavrakakis, and A. Ehrlich. *Discourse theory and political analysis: Identities, hegemonies and social change.* Manchester University Press, 2000.

[99] Z. Oravecz, K. Faust, W. H. Batchelder, and D. A. Levitis. Studying the existence and attributes of consensus on psychological concepts by a cognitive psychometric model. *The American journal of psychology*, 128(1):61–75, 2015.

[100] D. R. Oxley, K. B. Smith, J. R. Alford, M. V. Hibbing, J. L. Miller, M. Scalora, P. K. Hatemi, and J. R. Hibbing. Political attitudes vary with physiological traits. *science*, 321(5896):1667–1670, 2008.

[101] J. Park, S. Tauber, K. A. Jameson, and L. Narens. The evolution of shared concepts in changing populations. *Review of Philosophy and Psychology*, 10(3):479–498, 2019.

[102] P. Plato. Complete works, ed. john m. cooper, 1997.

[103] K. T. Poole and H. Rosenthal. *Ideology & congress: A political economic history of roll call voting.* Routledge, 2017.

[104] A. Puglisi, A. Baronchelli, and V. Loreto. Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, 105(23):7936–7940, 2008.

[105] R. A. Rappaport and R. A. R. Rappaport. *Ritual and Religion in the Making of Humanity*, volume 110. Cambridge University Press, 1999.

[106] T. Regier, P. Kay, and R. S. Cook. Focal colors are universal after all. *Proceedings of the National Academy of Sciences USA*, 102(23):8386–8391, 2005.

[107] T. Regier, P. Kay, and R. S. Cook. Universal foci and varying boundaries in linguistic color categories. In *Proceedings of the 27th Meeting of the Cognitive Science Society*, pages 1827–1832. Citeseer, 2005.

[108] T. Regier, P. Kay, and N. Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences USA*, 104(4):1436–1441, 2007.

[109] D. Roberson, J. Davidoff, I. R. Davies, and L. R. Shapiro. Colour categories and category acquisition in himba and english. *Progress in colour studies*, 2:159–172, 2006.

[110] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[111] B. A. Saunders and J. Van Brakel. Are there nontrivial constraints on colour categorization? *Behavioral and brain sciences*, 20(2):167–179, 1997.

[112] D. Schreiber, G. Fonzo, A. N. Simmons, C. T. Dawes, T. Flagan, J. H. Fowler, and M. P. Paulus. Red brain, blue brain: Evaluative processes differ in democrats and republicans. *PLoS one*, 8(2):e52970, 2013.

[113] C. G. Sibley and J. Duckitt. Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, 12(3):248–279, 2008.

[114] M. Singh and B. Arbad. Characterization of 4th–5th century ad earthen plaster support layers of ajanta mural paintings. *Construction and Building materials*, 82:142–154, 2015.

[115] R. Skotheim. The historian and the climate of opinion. 1969.

[116] B. Skryms. *Evolution of the social contract.* Cambridge University Press, 1996.

[117] K. Smith, S. Kirby, and H. Brighton. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386, 2003.

[118] P. Sorokin. Contemporary sociological theories. *American Journal of Sociology*, 34(2):382–384, 1928.

[119] R. Stark and R. Finke. *Acts of faith.* University of California press, 2000.

[120] L. Steels. Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308–312, 2003.

[121] L. Steels. Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4):339–356, 2011.

[122] Z. Sternhell, M. Sznajder, and M. Asheri. *The birth of fascist ideology: from cultural rebellion to political revolution.* Princeton University Press, 1994.

[123] W. G. Sumner. *Folkways: A study of the sociological importance of usages, manners, customs, mores, and morals.* Good Press, 2019.

[124] S. Tarrow. The language of contention: Revolutions in words. 2013.

[125] F. Tria, A. Mukherjee, A. Baronchelli, A. Puglisi, and V. Loreto. A fast no-rejection algorithm for the category game. *Journal of Computational Science*, 2(4):316–323, 2011.

[126] E. Trinkaus and P. Shipman. Neandertals: images of ourselves. *Evolutionary Anthropology: Issues, News, and Reviews*, 1(6):194–201, 1993.

[127] G. C. Vreugdenhil. Figures and tables. In *Psalm 91 and Demonic Menace*. Brill, 2020.

[128] G. Wallas. *The great society: A psychological analysis*. Macmillan, 1914.

[129] G. Wallas. *Human nature in politics*. Transaction Publishers, 1920.

[130] G. Wallas. *Our social heritage*. Yale University Press, 1921.

[131] B. Waubert de Puiseau, A. Afalg, E. Erdfelder, and D. M. Bernstein. Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of experimental psychology: applied*, 18(4):390, 2012.

[132] M. Weber. The protestant ethic and the spirit of capitalism. 1958.

[133] M. A. Webster and P. Kay. Individual and population differences in focal colors. *The anthropology of color*, 2007.

[134] O. D. Weeks. Mcgovern, william m., from luther to hitler; the history of fascist-nazi political philosophy.(book review). *Social Science Quarterly*, 23:184, 1942.

[135] S. C. Weller, R. D. Baer, J. G. de Alba Garcia, and A. L. S. Rocha. Explanatory models of diabetes in the us and mexico: The patient–provider gap and cultural competence. *Social Science & Medicine*, 75(6):1088–1096, 2012.

[136] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

Appendix A

Schematic Similarity: Metric Proof

In mathematics, a *metric* is defined as a function that describes the distance between pairs of elements within a set. A metric $d(x, y)$ defined on the set S also satisfies the following axioms, such that $\forall x, y, z \in S$:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \iff x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

Our measure, called *Schematic Similarity* (SS), describes an inverse distance relationship between pairs of elements (i.e. WCS participants). Therefore, we let $d(x, y) = 1 - SS(x, y)$ be our distance function. Here we provide a formal proof, organized according to the four axioms listed above, that $d(x, y)$ is a mathematical metric.

Theorem A.1. *The distance function $d(x, y) = 1 - SS(x, y)$ defined on the set X, which is the set of all WCS participants, is a metric.*

Proof. Let $x, y, z \in X$. WLOG, assume that x is the *Reference Participant* (i.e. participant using fewer color terms) and y is the *Other Participant.*

Axiom 1: d(x,y) $\geq$ 0

Let M be the set of mapped terms between participants x and y. (See Section 2.3.4 for more information on how to generate this set.)

By definition, $d(x, y) = 1 - SS(x, y)$.

Since $0 \leq SS(x, y) \leq 1\ \forall x, y \in X$, then $d(x, y) \geq 0\ \forall x, y \in X$.

Axiom 2: d(x,y)=0 $\iff$ x=y

($\Longleftarrow$) Assume $x = y$ (i.e. x and y are the same participant).

Then the set of mapped terms is given by:

$$M = \bigcup_{i=1}^{n} \{(i, i)\}$$

where $n =$ number of terms x used and each i represents an individual term.

Since $x = y$, $\forall (i, j) \in M$, $i = j$.

Therefore, $|T_i^x \cap T_i^x| = |T_i^x \cup T_i^x|$ and $e_{T_i^x} = 0$ because $U = \emptyset$.

Hence,

$$\begin{aligned}
d(x,x) &= 1 - SS(x,x)\\
&= 1 - \sum_{(i,i)\in M} \frac{1}{n} \frac{|T_i^x \cap T_i^x|}{|T_i^x \cup T_i^x| + e_{T_i^x}}\\
&= 1 - \sum_{(i,i)\in M} \frac{1}{n} \frac{|T_i^x \cap T_i^x|}{|T_i^x \cup T_i^x|}\\
&= 1 - \sum_{(i,i)\in M} \frac{1}{n} * 1\\
&= 1 - \frac{n}{n}\\
&= 1 - 1\\
&= 0
\end{aligned}$$

($\Longrightarrow$) Assume that $d(x,y) = 0$. Then,

$$\begin{aligned}
1 - SS(x,y) &= 0\\
SS(x,y) &= 1\\
\sum_{(i,j)\in M} \frac{1}{n} \frac{|T_i^x \cap T_j^y|}{|T_i^x \cup T_j^y| + e_{T_i^x}} &= 1\\
\sum_{(i,j)\in M} \frac{|T_i^x \cap T_j^y|}{|T_i^x \cup T_j^y| + e_{T_i^x}} &= n
\end{aligned}$$

Since $0 \leq SS(x,y) \leq 1$ and $|M| = n$, then

$$\frac{|T_i^x \cap T_j^y|}{|T_i^x \cup T_j^y| + e_{T_i^x}} = 1, \qquad \forall (i,j) \in M$$

Therefore, $|T_i^x \cap T_j^y| = |T_i^x \cup T_j^y| + e_{T_i^x}$.

By construction, $e_{T_i^x} \in Z^+ \cup \{0\}$.

Suppose $e_{T_i^x} > 0$.

Then $|T_i^x \cap T_j^y| > |T_i^x \cup T_j^y|$, which is a contradiction.

Therefore, $e_{T_i^x} = 0$.

It then follows that $|T_i^x \cap T_j^y| = |T_i^x \cup T_j^y|$. In other words, for every term i used by participant x and its mapped term j used by participant y, the set of chips being called i and j by x and y, respectively, are exactly equal. That is, terms i and j refer to the exact same regions of the color space.

Therefore, the mapping from terms used by x and terms used by y is a bijection.

Since x and y use the same number of color terms and each mapped pair of terms refers to the exact same region of the color space, participants x and y have identical partitions of the color space. Therefore, $x = y$.

Axiom 3: d(x,y) = d(y,x)

Recall that x is assumed to be the *Reference Participant (RP)* and y is assumed to be the *Other Participant* (OP).

Then $SS(x, y) = SS(y, x)$ because *Schematic Similarity* only cares about who is the RP and who is the OP, not about the order in which they are given to the function. Therefore,

$$\begin{aligned} 1 - d(x, y) = SS(x, y) = SS(y, x) = 1 - d(y, x) \\ 1 - d(x, y) = 1 - d(y, x) \\ d(x, y) = d(y, x) \end{aligned}$$

Axiom 4: d(x,z) ≤ d(x,y) + d(y,z)

Due to the formulation of *Schematic Similarity*, it is impossible to prove the Triangle Inequality of our distance metric analytically. Therefore, we computed $d(x, y)$, $d(y, z)$, and $d(x, z)$ $\forall (x, y, z) \in X \times X \times X$ and checked $d(x, z) \leq d(x, y) + d(y, z)$. We found this inequality to hold for every combination of WCS participants. For our purposes, this is sufficient evidence that the Triangle Inequality holds for our distance metric d. □

Appendix B

Chapter 3 Clustering Alpha distribution

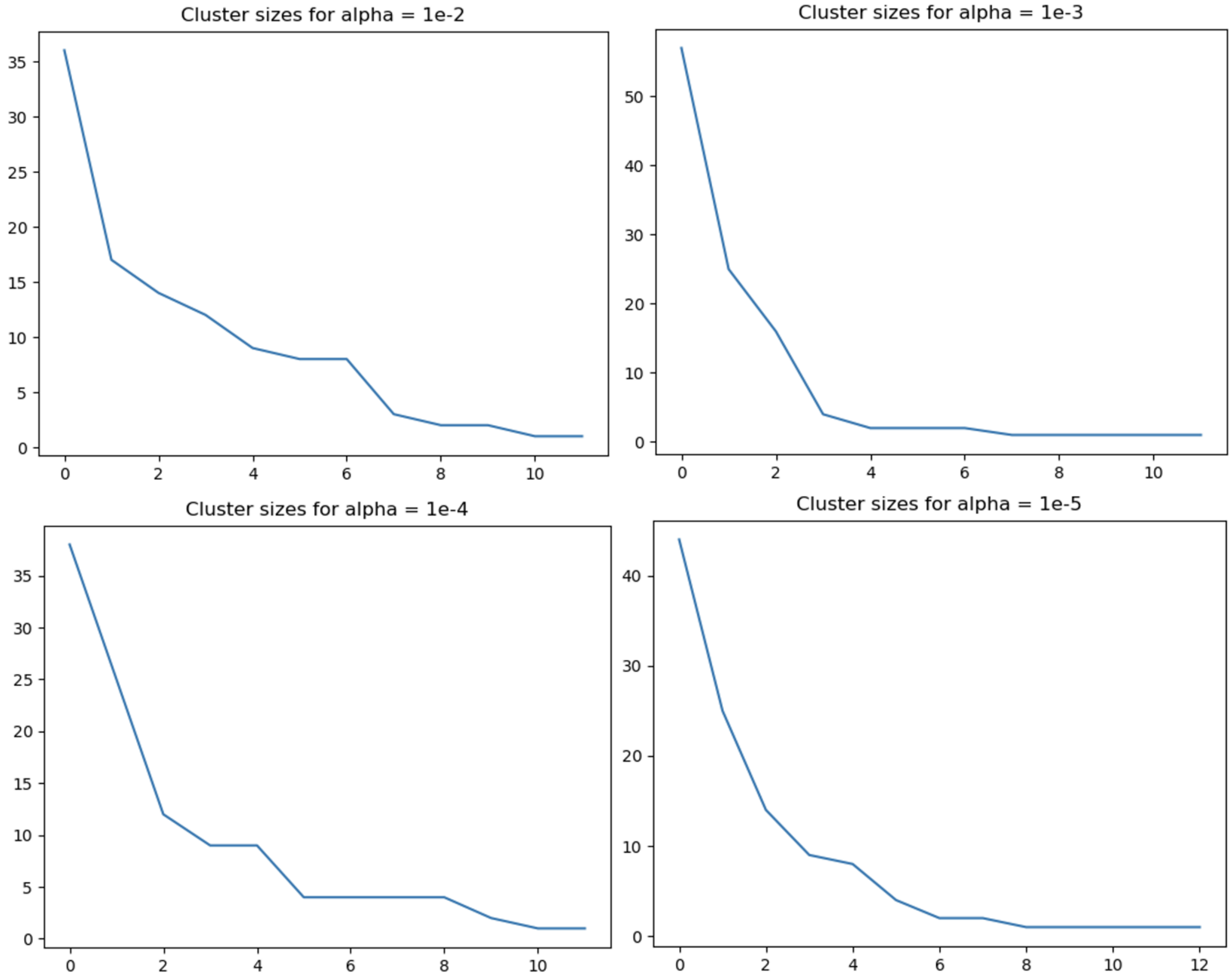


Figure B.1: The relationship the Dirichlet hyper parameter α and size of cluster.