June 2022

Interdisciplinary Communication by Plausible Analogies: the Case of Buddhism and Artificial Intelligence

Michael Cooper
University of South Florida

Interdisciplinary Communication by Plausible Analogies: the Case of Buddhism and Artificial Intelligence

by

Michael Cooper

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Philosophy
College of Arts and Sciences
University of South Florida

Co-Major Professor: Stephen Turner, Ph.D.
Co-Major Professor: Wei Zhang, Ph.D.
John Licato, Ph.D.
Alex Levine, Ph.D.
William Goodwin, M.S.M.E.

Date of Approval:
March 22, 2022

Keywords: Buddhist Philosophy, Process Metaphysics, Analogy, Kuhn, Artificial Intelligence

Table of Contents

List of Figures

Abstract

Communicating interdisciplinary information is difficult, even when two fields are ostensibly discussing the same topic. In this work, I'll discuss the capacity for analogical reasoning to provide a framework for developing novel judgments utilizing similarities in separate domains. I argue that analogies are best modeled after Paul Bartha's By Parallel Reasoning, and that they can be used to create a Toulmin-style warrant that expresses a generalization. I argue that these comparisons provide insights into interdisciplinary research. In order to demonstrate this concept, I will demonstrate that fruitful comparisons can be made between Buddhism and Artificial Intelligence research.

Chapter 1: Introduction

Interdisciplinary science is controversial. Making the theories of one discipline relevant to research done in another is a difficult task, but it offers significant rewards. Many funding agencies imply that interdisciplinary research is especially to be sought after, since it's arguably more likely to produce revolutionary leaps forward than mere evolutionary advances in a discipline. In order to do interdisciplinary science well, or to criticize it when it goes wrong, it's useful to have a clear model of how information from one field can become relevant to an entirely different field. In this dissertation, I articulate a model of analogical reasoning that could facilitate interdisciplinary advances. In order to do so, I'll dive into the problem of incommensurability, an important issue in this area.

1.1 Kuhn and Analogies

A common criticism of interdisciplinary work is that the situations in the two domains is incommensurable, to use Thomas Kuhn's language. To say that they are incommensurable is to say that the two domains that we're trying to compare are so different that they cannot be compared without some loss in detail, due to fundamental differences in the two domains which prevent lossless communication. This interpretation of incommensurability lies in contrast to a more restrictive view of incommensurability which indicates that it is impossible for two separate domains to usefully interact. To begin addressing this problem, I'll begin by asking whether the charge of incommensurability applies to this situation at all.

Kuhn's arguments about incommensurability are fundamentally questioning whether a devotee of a scientific paradigm can effectively communicate with a researcher committed to another scientific paradigm, particularly if the second paradigm threatens to overturn

the first. There are two ways to see this as applying to the issue of this dissertation. In the first place, we can interpret Buddhist philosophy as representing a paradigm, or set of paradigms, which I'm proposing to compare to scientific paradigms in Cognitive Science. In the second place, if it is applied here, incommensurability is a challenge to the effectiveness of scientific analogies altogether, since analogies are essentially about a comparison between two (often very different) domains. These domains, however, do not have to conform to what Kuhn considers to be a "paradigm." In Bartha's literature review, he notes that some of the considerations of Kuhn's paradigms seem to apply to arguments by analogy, but not others:

> This role of a metaphor in the "logic" of science, [Stepan 1996] suggests, comes close to that of a Kuhnian paradigm. Such metaphors may prove to be helpful or they may (as in her example) prove to be harmful, but in any case they are indispensable... It is important here to distinguish between the broad role of analogy, in providing an underlying metaphor for a sustained research program, and the narrow role of grounding an individual analogical argument... the ideas Kuhn introduced in The Structure of Scientific Revolutions help to make this distinction clear. An underlying metaphor can function, like a paradigm, as a constitutive element in scientific work. Metaphors can provide models and images that shape our perception of phenomena, and they can establish a grip that is difficult to shake. Individual analogical arguments, by contrast, are less grand. In Kuhnian terms, most of them extend an existing paradigm; they are employed in the context of normal science.(Bartha, 2010, p.11-2)

Though a grand metaphor might provide the basis for a paradigm to establish itself, the metaphor does not appear to be identical to the paradigm. An analogy, of the kind I'm describing in this work, is similar to a metaphor, but unlike a metaphor it is simplified and idealized in a structure. This is hardly a surprising result, as Masterman (1970) showed that there were arguably dozens of different definitions of "paradigm" in Kuhn's Kuhn (1962) work, but that analogical reasoning is consistent with some of those definitions. In this

work, I'll avoid some of the complexities of the multifariously defined word. Instead, I'll focus on arguments by analogy, as I've defined them here, with the understanding that only some of Kuhn's worries about paradigms apply to this area.

Kuhn primarily discussed incommensurability in reference to people using different uncontroversially scientific paradigms, especially in the case of pre- and post-revolutionary scientists speaking with one another, but incommensurability's lessons apply more generally. In The Road Since Structure, Kuhn gives some explicit formulations of what he means by the term: "The claim that two theories are incommensurable is then the claim that there is no language, neutral or otherwise, into which both theories, conceived as sets of sentences, can be translated without residue or loss." (Kuhn and Wilson, 2001, p.36) The criticism that Buddhism and Science are incommensurable, under this formulation, is essentially that the two can't speak fruitfully to one another. At first glance this seems to argue that even when it appears that the two disciplines are using the same terms, those terms might have subtle definitional differences that become obstacles to effective communication. To claim, then, that the two are incommensurable is to say that even were the disciplines to engage with each other more, we should expect the efforts to make little progress because the two sides will be "talking past each other."

The problem with taking incommensurability at first glance in that way, though, is that this would make effective communication between any two disciplines which use different theories and methods impossible, which clearly doesn't seem to be the case. The rarity of scientific revolutions might make it plausible at first glance that the people using the old paradigm can't effectively communicate with those using the new paradigm, but there is a commoner situation of different paradigms in different contemporaneous fields which casts doubt on that. Researchers from related scientific disciplines (or sub-disciplines) clearly seem to have different paradigms and manage to communicate effectively. Kuhn is well aware of this objection in The Road Since Structure and points out that the objector has made the incommensurability argument too strong in order to defeat it. He points out that there is

an important distinction to be made between incommensurability and incommunicability. If two things are incommensurable, they can't be directly compared without loss because between them there is "..no common measure. But lack of a common measure does not make comparison impossible. On the contrary, incommensurable magnitudes can be compared to any required degree of approximation." (Kuhn and Wilson, 2001, 35) So Kuhn is not arguing that researchers from two disciplines with different paradigms can't speak to each other at all. It's just that researchers from different disciplines can't communicate without some loss of information along the way, related to the relevant differences between their disciplines. Applied to Buddhism's theories of mind and neuroscience, the question becomes, "Is there so much lost in translation between Buddhism and neuroscience that the effort is not worth the potential rewards?"

1.2 Wallace and Gould Claim it is Impossible

The difficulties of communication between disciplines depends on how much the disciplines have in common, so it makes sense to examine the overlap. At the extreme end, many people argue that religion and science have entirely non-overlapping domains. Since the result of Kuhnian incommensurability is consistent with more traditional arguments about the incommensurability between science and religion, that lends support for the decision to extend Kuhnian reasoning beyond the bounds of scientific disciplines. Paleontologist Stephen Jay Gould, in his book Rocks of Ages: Science and Religion in the Fullness of Life made a famous argument along these lines. He wrote that science and religion operate in non-overlapping magisteria that nonetheless need to be integrated in some way to lead a full life. To make this case, he argued that science concerns empirical data, and religion, by contrast, concerns human purposes, meaning, and value. "...the net, or magisterium, of science covers the empirical realm: what is the universe made of (fact) and why does it work this way (theory). The magisterium of religion extends over questions of ultimate meaning and moral value. These two magisteria do not overlap..." (Gould, 2007, 10) This

notion has some initial plausibility, as it is arguably impossible to start with either epistemic information or information about human purposes, meaning, and value and derive the other kind of information from it.

However, once this argument is examined problems quickly emerge. Though you can't derive a value from epistemic information, epistemic information can be essential to matters of value. For example, if people determine that keeping the mean global temperature stable is a morally desirable state, it then becomes vitally important for them to understand the extremely difficult epistemic problem of understanding any effects of our actions on the mean global temperature. Values can also be essential for determining scientific practice. It would be impossible to determine how best to test new drugs, for example, without a careful balancing of values: potential benefits to patients must be weighed against unknown risks and constrained by the rights of the test subjects. Gould tries to address these kinds of difficulties by making science and religion adjacent to one another, even if they don't overlap.

> I hold that this non-overlapping runs to completion only in the important logical sense that standards for legitimate questions, and criteria for resolution, force the magisteria apart on the model of immiscibility—the oil and water of a common metaphorical image. But, like those layers of oil and water once again, the contact between magisteria could not be more intimate and pressing over every square micrometer (or upon every jot and tittle, to use an image from the other magisterium) of contact. (Gould, 2007, 32)

Gould bases his assertion of the non-overlapping quality of the magisteria on the idea that they concern separate subject matter and therefore have different criteria for questioning and for answering those questions. It is in this sense that he thinks the two are immiscible. According to this argument, there are three areas in which Buddhism and science should be analyzed, to see if they can be usefully mixed. The first concerns whether Buddhism is a religion in the sense that Gould uses it here. The second concerns whether the subject matter

of Buddhism and science overlap, despite Gould's objection. The third concerns whether Buddhism and science use relevantly different criteria for asking and answering questions. In any of those three areas, there are plausible concerns which might make Buddhist theories of mind and neuroscience immiscible.

The argument so far has been that the claim of Kuhnean incommensurability rests on the idea that something is lost in translation during the interaction. The interactions that Buddhist theories of mind and neurologists are likely to have will begin with defining the detectable physical correlates of states of mind generated by Buddhist practitioners. An early example of this has been provided by the psychologists Richard Davidson and Antoine Lutz in their EEG and fMRI study of Matthieu Ricard. Practically speaking, the question is whether scientists and Buddhist practitioners can effectively communicate about the results of experiments like this without something vital being lost in translation. Do the different systems of Buddhist practice and neurology differ in some essential respects to the point where they can't speak effectively to each other? Jay Gould might argue that if Buddhism belongs in the category of "religion," then their subject matters are too distinct from one another for their communication to be accurate to the required degree. The label doesn't seem to fit Buddhism very well, though.

Defining the word "religion" is a notoriously difficult task, at least if we want a definition that meets the standards usually demanded of philosophical definitions. It would be nice to have some necessary and sufficient set of conditions that all and only religions have. The ideal definition would also include all of the various doctrines we would normally consider religions, and settle edge cases like Confucianism. Unfortunately, no such definition seems likely to work. As Alan Wallace argues in Buddhism and Science, "To understand Buddhism on its own terms, it is imperative that we in the West recognize the cultural specificity of our own terms religion, philosophy, and science and not assume from the outset that Buddhism will somehow naturally conform to our linguistic categories and ideological assumptions." (Wallace, 2007, 5) Instead of looking for necessary and sufficient conditions, it may be

possible to treat the extension of the word "religion" as a case of Wittgensteinian family resemblance. Our understanding and definition of religion in the west can be most charitably understood as based on the paradigm case of Judeo-Christian experience and extended from there to similar cases, but this raises problems. It arguably makes our understanding of the word "religion" ethnocentric, because we're defining a supposedly objective category on the basis of local culture.

Another problem with this Wittgensteinian definition for religion is that it doesn't provide certainty on the question of group membership. If being a religion is simply similarity with accepted paradigm cases of religion, we'll have to deal with the fact that Buddhism is different from the paradigm cases in numerous ways and it's unclear if those differences put it outside of the scope of the term. This failure of categorization is itself worrying, as membership in such categories heavily influences disciplinary relationships. As Wallace argues

> This same line of reasoning is the one used for excluding Buddhist philosophy from virtually all academic departments of philosophy in Europe and America: if Buddhists don't philosophize following the same rules as Western philosophers, they don't philosophize at all. But if we should follow this line of reasoning ad absurdam, since Buddhism does not even affirm the existence of a divine Creator who rules the universe, punishes sinners, and rewards the faithful, like the "genuine" religion of Christianity, it can't even be counted as a religion. It simply falls through the cracks and counts for nothing at all. (Wallace, 2007, p.6)

It seems that the inevitable result of badly categorizing Buddhism entails taking it out of communication with some other discipline. The category of "religion" is too poorly defined for such an important determination to turn on it. A closer look at the features of Buddhism is warranted, to see what it shares with science and where it seems to conflict.

1.3 Are the Dissimilarities Unconquerable?

Some elements of Buddhism are straightforwardly metaphysical and arguably incommensurable with science as in the continuity between one lifetime and the next in reincarnation. Buddhists believe in reincarnation, and justify that belief by pointing to examples of children that appear to remember verifiable information about the lives of dead people. People outside of the faith are reluctant to accept this information, though, given the lack of a verifiable causal mechanism for the transfer of memories past death. Buddhists also believe that through specific meditative practices, it's possible for well-trained people to gain accurate knowledge of the workings of karma, envisioned as a law of nature which causes deserved punishments and rewards. This also fails to gain acceptance outside of Buddhism due to disagreements about the 12 proposed causal links which purport to explain this relationship. Although these practices are certainly incompatible with accepted scientific standards for evidence, these are not the central focus of Buddhism. I see no reason why Buddhism can't be understood piecemeal, with some elements commensurate with science and other elements not, since the issue in question is the degree of overlap. I propose, therefore, to move past cases which are clearly inconsistent with scientific practice to examine cases that might be consistent.

There are a number of features Buddhism may have in common with science. This is important because Gould's argument indicates that science and religion occupy non-overlapping magisteria, even if he does hedge by arguing that they are in close connection with each other. It's enough to counter his argument, therefore, by showing that many areas do in fact overlap. In Thupten Jinpa's chapter in Wallace's Buddhism and Science he argues that there are a lot of features in Buddhism which are arguably consistent with the subjects and goals of science:

> It may well be that of all religions Buddhism finds it easiest to engage in a critical dialogue with science. The following key features of Buddhism— its suspicion

> of any notion of absolutes, its insistence on belief based on understanding, its empiricist philosophical orientation, its minute analysis of the nature of mind and its various modalities, and its overwhelming emphasis on knowledge gained through personal experience—all make it easy for Buddhism to be in a dialogue with a system of thought that emphasizes empirical evidence as the key means of acquiring knowledge. (Wallace, 2007, 83)

The features that Jinpa identifies here suggest a number of areas in which Buddhism (including theories of mind) and science are speaking about similar topics and using consistent methods. In this work, I argue that, though there are relevant differences between the subject matters of Buddhist philosophy and theories in cognitive science (especially in the area of natural language processing in artificial intelligence research. One likely objection to the compatibility of Buddhist theories of mind and neuroscience is to argue that science is an entirely naturalized enterprise, while Buddhism is concerned with the supernatural. Owen Flanagan, in his book Buddhism Naturalized, provides a useful explanation of naturalism in this context:

Naturalism comes in many varieties, but the entry-level union card—David Hume is our hero—expresses solidarity with this motto: "Just say no to the supernatural." Rebirths, heavens, hells, creator gods, teams of gods, village demons, miracles, divine retributions in the form of plagues, earthquakes, tsunamis are things naturalists don't believe in. What there is, and all there is, is natural stuff, and everything that happens has some set of natural causes that produce it—although we may not be able to figure out what these causes are or were. (Flanagan, 2011, 2)

By this definition, naturalism seems focused on the non-existence of supernatural causes. That might suggest that we need to catalogue all of the causes to discover the lack of any cause which extends beyond material reality. However, t's not necessary for us to understand the causes of everything in the universe to be naturalists as the expectation that such causes could be found is sufficient.

While there are certainly elements of Buddhism that a scientist would likely consider supernatural, that's not enough to dismiss Buddhist theories of mind from consideration, given the piecemeal approach I am describing. Does the worry that Buddhism is concerned with the supernatural extend to the point that they cause worries about our theories of mind? I argue that the problem doesn't extend that far because it's hard to link theories of mind to substances, whether those substances are conceived of as natural or supernatural. Criticisms of the compatibility of Buddhist theories of mind and neuroscience could argue that Buddhist psychology is not sufficiently materialist, arguing that it concerns elements of mind that may not be made of physical matter. This criticism would be tantamount to reductionism, since they insist that we be able to trace theories of mind back to a specific kind (natural) of substance. Reductionism has some well-known flaws, though. As Wallace explains, it is not a very good fit for the study of the mind:

> In the brain sciences, for example, if one focuses one's attention on the operations of individual subatomic particles, atoms, molecules, cells, or even entire ganglia of neurons, this excessively narrow vision can obstruct insight into the global processes occurring in diverse regions of the brain. Moreover, if one focuses solely on objective brain functions and ignores subjective mental events, this mode of reductionism prevents one from discovering mind-brain correlates. (Wallace, 2007, 12)

Reductionism is a bad standard to use, since much of brain science can't be meaningfully attempted on the basis of the physics of neurons. We simply do not have instruments fine enough to trace every mental state back to base physics, even if we had adequate theories to do the interpreting. Furthermore, if reductionism was true, then reductionist criticisms of Buddhist theories of mind would apply to western psychology as well. Like it or not, mental theories of any kind will have to be discussed without tracing everything back to material correlates for the time being. That being the case, what is the point in insisting that they all trace back to natural things? Nothing of relevance to making improvements

to our theories of mind seems to turn on the question, as long as rigorous methods are used that don't assume supernatural causation or substances on the basis of faith. Since no one I'm aware of is proposing that scientists take supernatural causation or substances on faith, that doesn't seem to be a significant problem.

As the above discussion reveals, accepting that Buddhist theories of mind can fruitfully speak to neurological worries depends on whether Buddhist reasoning uses scientifically acceptable causal stories. This opens up potential worries that Buddhism isn't telling those kinds of causal stories. These worries are amplified when we consider that Buddhist theories of the mind tend to focus on the phenomenology of experience, rather than focusing on physical material. This is less worrying, though, when we reject reductionism and observe that neurology is also concerned with phenomenological experience. Both disciplines seem to overlap in their subject matter in this regard, casting some doubt on Gould's assertion of non-overlapping magisteria.

There remains a worry, though, that even if they overlap in subject matter, they may not overlap in methods. Wallace makes the case that the methods of Buddhist psychology are arguably consistent with science in the rigor of their established methods:

> ...it has also, from its very origins, established rigorous methods for experientially exploring the personal and impersonal phenomena that make up the natural world. Such techniques, many of which are designated by the English term meditation, frequently entail careful observation followed by rational analysis. In short, there are elements of Buddhist theory and practice that may be deemed scientific. (Wallace, 2007, p.5)

In Kuhnean terms, someone might worry that Buddhist theories of mind are in a pre-paradigm state, so that practitioners need to reinvent their discipline in every paper in order to argue for an advance. That doesn't appear to be a concern for Buddhism, since its methods are long-established and have already gained the acceptance of a significant body

of practitioners. This creates a framework for a Buddhist study of mind which is capable of puzzle-solving and making advances on the basis of new observations.

Another worry is that Buddhist theories of mind rely on private evidence, which can't be scientifically assessed. This is a significant worry, as scientific results need to be publicly observable so that other practitioners can attempt the same experiments and accurately judge whether they have achieved the same results. However, the worry about private evidence is unfounded, since the established practice of collecting and evaluating introspective observations that result from Buddhist introspection has a long history of evaluation by experts using consistent theories. The results of Buddhist meditative experiments are certainly not capable of being assessed by the average person, but as Wallace points out, that's an unreasonably strict requirement.

> ...even after it is published a scientific discovery can normally be validated only by a relatively small number of experts within a specific field of research. Other scientists and the general public will, for the most part, accept the discovery on the basis of their faith in the experts. This situation is not so different from discoveries made by Buddhist contemplatives. (Wallace, 2007, p.9)

The results of science need to be publicly accessible, but training in the paradigms of science is necessary to interpret its results. The same considerations can be applied to experiments within Buddhist theories of mind. Buddhist methods of introspection produce insights as part of a expert-guided training of the mind, which aims, among other things, to clarify our understanding of the world. The individual introspective insights that result from this practice are predictable according to the established body of knowledge, which is preserved in canonical texts, well-regarded commentary literature, and training manuals. These sources establish the theories under which introspective insights are to be elicited and evaluated. Far from being mere private evidence, Buddhist methods represent established theories which might be consulted for insights into how the mind works, even under strict Kuhnean consideration. Further, the results of these practices can be made more scientific,

following the model of western psychology, by relying on surveys of statistically significant numbers of people practicing Buddhist methods. It seems that there is plenty of overlap between the subject matter and methods of science and those of Buddhist theories of mind to avoid Gould's conclusion. Another worry remains, however, that the values present in Buddhism make it incommensurable with science. If non-shared values pervade the area of the perceived overlap, perhaps that will make terms in Buddhism incommensurable with terms in neurology.

S

In this work, I take a pragmatic approach to scientific theories. Under this approach, a good scientific model is one that approximates the truth with fidelity that improves over time, but is never absolute. Given this outlook, it's not appropriate to say that a scientific model matches reality, but instead that it is relevantly similar or dissimilar to the situation it is modeled upon. Analogical reasoning is particularly useful for closely analyzing the similarities and dissimilarities of complex and dynamic systems and provides a concrete framework from which to analyze relevance, a necessary part of the task.

By mostly leaving AI research out of his accounts, Wallace blinds himself to the important information that might be gleaned about the functional ways we think. Instead, he imagines that these efforts will only be worthwhile if they yield physical information about the tiniest parts of the brain. In this quote, we see that Wallace disregards the value of understanding the functional elements of the mind, implying that such research will only have value if it speaks to the physical components of the brain.

> Functionalist accounts have been very prevalent in these recent brain-centered theories of the mind, but it is not clear what, if any, information they provide as to the real nature of what humans do when we introspect... if introspection in this sense were to provide us with immediate access to and knowledge of the brain, it would yield knowledge about neuronal firings, the state of the neuron-

> protecting glial cells, and the intricacies of cerebral processes and states; but this has not proven to be the case. (Wallace, 2004, p. 78)

Functionalist accounts of the mind in computer science are used to describe the behaviors of the mind in such broad terms that those functions might be replicated by other means. When a program like GPT-3 has success in creating human-like responses to queries, it reveals something about which functional relationships are possible and successful. This is valuable even if it is never translated into insights into the material nature of the mind.

> Religious fundamentalists regard those who reject their dogma as being victims of their own sin, especially the sin of pride. Similarly, champions of scientism condemn dissenters from their view as having abandoned reason, for it is inconceivable to them that anyone could be rational and knowledgeable of science yet deny their most cherished scientific beliefs. In short, scientism is to scientific materialism what fundamentalism is to all traditional religions.(Wallace, 2004, p. 31)

In this statement of a fundamental problem for contemplative science, I'm in agreement with Wallace. The problem is, his rejection of agnosticism leaves him no other alternatives than to fall on one or the other horn of this dilemma. In injecting religious ethical and epistemological principles, he falls onto the horn of the dogmatic religionist.

Given his assessment of the problem, it makes sense that he thinks the solution is to get scientists to reject materialism in favor of Buddhist metaphysics. He argues that scientists are stuck in improper theories which are based, among other things, in scholastic philosophy, and that this has subtly introduced western religious metaphysics into secular science. Science is tainted by the metaphysical assumptions in its founding and development, which linger still.

> According to a common materialistic viewpoint, human beings are identical to our brains and all our activities are governed by the laws of physics, so the experience of choosing is an illusion. Given the limitations of the current scientific

> understanding of consciousness, these assertions are simply beliefs, determined in large part by inductive reasoning based on nineteenth-century materialism. Current empirical evidence and rational analysis do not compel anyone to accept these statements; those who have adopted them have chosen to do so, although they may feel they have no alternative. (Wallace, 2013, 10)

He assumes that these theories are still in place, perhaps unconsciously applying a religious model. However, religious theories harken back to original founders in a way that scientific theories usually do not. New scientific theories are sometimes meant to supplant old ones entirely. The views of the founders of any scientific branch are a matter for historical scholarship, not current practice. Recall that his earlier rejection of Agnosticism was to make it simply another kind of religion. Wallace thinks that once Buddhist metaphysics and methods are credited and used, then Contemplative Science can proceed. In the next section, I'll explain why I think this view is mistaken.

Chapter 2: Buddhist Philosophy of Mind and AI

2.1 The Buddhist view of Cognition

Buddhism is concerned with fostering positive, healthy, and beneficial emotional health. In order to understand this process, its doctrine carefully analyzes mental functions. It is particularly concerned with the processing raw sensory input into recognized objects of consciousness. "mental cognitive awareness (manovinnana, often translated literally as 'mind-consciousness')... manovinnana refers to the distinctive awareness that is the cognitive basis of sense perception issuing from the contact between manodhatu and its respective dhamma objects." (Ronkin, 2005, p. 39) It's important to note that this analysis aims to explain the different mental functions that we experience and how they depend on one another. It does not aim to describe the (material or immaterial) substances that make up the mind.

This effort to promote intellectual well-being is intended to help practitioners escape the cycle of rebirth (*Samsara*) and reach Nirvana. Though this is a soteriological goal in Buddhism, the desire to address mental suffering is also the core normative principle of psychology. As I argued in chapter 1, the normativity involved in avoiding suffering is not one that disqualifies a practice from scientific scrutiny, as long as the goal of seeing the world accurately is also respected. The 14th Dalai Lama writes about this central focus in Wallace's Buddhism and Science

> Specifically, it is important to recognize which mental processes, especially emotions, are incompatible with each other. Moreover, it is crucial to investigate with discerning intelligence which emotions are truly beneficial over the long run and which are harmful. In conjunction with that, one should study which emotions

> are in accord with reality and which are misleading. Given the importance of understanding this, it is apparent that one also needs to gain a precise understanding of the objects apprehended by the mind. This leads one to investigate whether an object that appears to the mind actually exists in accordance with the way it appears. (Wallace, 2007, p.101)

In order for Buddhists to achieve their goal of allieviating suffering, according to this doctrine, it's necessary to begin with a detailed assessment of mental functions, with an eye towards making sure our conceptions of match with the world. As Ronkin argues, "This theory portrays each moment of discriminative, cognitive awareness as involving not only the occurrence of sense perception itself, but also the derivation of a series of other related mental events. For instance, visual perception involves not only seeing itself, but such occurrences as adverting to the appropriate sense 'door', fixing of the visual object in the mind, examination and recognition of its features and identification of its nature..." (Ronkin, 2005, p. 40) Given this focus on recognizing and categorizing the world in an accurate way, I argue that the normative goals described in that project are compatible with the goals of psychology for improved mental health. Though it would be difficult to define the normativity of this drive with scientifically-acceptable precision, there are surely some aspects of this discipline which are amenable to testing.

2.2 No Self

One doctrine which is of central importance in Buddhist philosophy of mind is the doctrine of no self (*anatta.*) The doctrine is a denial of many of the common ways that we interpret the notion of the self, identifying the idea of a persistent self as a conventionally-useful illusion. It argues that, although we have a common convention of discussing the self as something that persists through time, unchanging and grounding personal identity, that notion is misleading. Instead, it argues that persons exist, in the sense that they are as a dynamic aggregation of mental functions. "The Buddhist strategy for overcoming this

mistake begins by distinguishing between two possible referents for the 'I' of the 'I'-sense: a self, understood as the one part among all the psychophysical elements that grounds diachronic personal identity; and a person, understood as the whole that is composed of the many psychophysical elements." (Siderits, 2020, p. 104) As Siderits argues, the denial of the self is meant to be a denial of an unchanging core of self. By contrast, Buddhism embraces a notion of self which is understood as an aggregate of processes subject to constant change.

While this foundational argument is often cited in ethical contexts, where a denial of the self is used to argue against self-centeredness, [Dambrun, 2012] it's also an essential doctrine for understanding the composition of Buddhism's model of mental functions. A principle difficulty in discussing the doctrine of no self is wondering how cognition, recognition, and memory could make sense without relying on a core, changeless notion of self for reference. This was one of the principle objections that was raised by scholars of the Hindu Nyaya school. The argument, as it applies to memory, worries that in order to remember, we need a consistent idea across time of the one who remembers. Vasubandhu explains this difficulty in terms of a connected series of causal interactions that result in the behaviors associated with having a memory: "It was said that what does that is the distinct cognition that is the cause of the memory. What is then expressed as 'Caitra remembers' is so called having perceived that [this distinct cognition] occurs due to the causal series called 'Caitra', it is thus said, 'Caitra remembers'." (Siderits, 2007, p. 125) Siderits suggests that this notion of time might be understood as if the memory is an odd kind of metaphorical seed, whose effects will only be felt later. The metaphorical seed lacks persistance through time, despite it's delayed effects. It causes it's successor seed until such time that one of the successor seeds sprouts, providing conscious access to the contents of that seed at a time far removed from the creation of the original seed.

One of the reasons that Buddhists argue that there is no self is a metaphysical view about the non-reality of aggregate objects. They argue against the position sometimes called "extreme realism" which dictates that the objects we perceive have a wholeness and

distinctness which is based on the object's nature. On the contrary, Buddhists argue that when we perceive an aggregate entity, like a heap of sand, the "heap" is a conceptual fiction which is established in our minds, not something that can be found in the world.

Vasubandhu approaches this problem with the evocative example of a piece of cloth, demonstrating that "cloth" is a mere conceptual fiction we create from our incomplete impressions of the aggregate arrangement of a group of threads. He does this in search of a core of the self that the Indian Vaiśeṣika metaphysicians (secular philosopher's whose beliefs were similar to the Hindu Nyāya school) supposed must exist. Vasubandhu argued instead that this essence wasn't a feature of the outside world, but instead a conceptual fiction which lives within our minds. This, seemingly abstract debate, was in service of explaining the nature of the 'self' we perceive and refer to in our grammars. He argued that, like the cloth, our mental cognitions are an aggregate of different senses and faculties which we only conventionally perceive as constituting a unified self.

In this first section of the extended passage, Vasubandhu examines the way we instinctually cognize objects as wholes which we only experience as separate parts through our senses. He does this in search of a central essence which gives an object its distinct nature. In examining the parts of a piece of cloth, he fails to find any such essence.

> Vaiśeṣika: - How do you establish that cloth is not a substance distinct from threads (dravyāntạra)?
>
> The Buddhist: - When the sense-faculty (of the eye and of the body) is in touch with one thread, the cloth is not perceived. Now, if the (whole) cloth existed (in each thread,) why would it not be perceived?
>
> If the Vaiśeṣika would say that the whole cloth does not exist in each thread (akṛtvavṛtti), then this is to acknowledge that the cloth is just the collection of its parts which are each constituted by one thread: for how would you prove that the parts of a cloth are something other than the threads?

> If the Vaiśeṣika would say that the whole cloth exists in each thread but one does not perceive the whole cloth in each thread because the perception of the cloth presupposes a connection of the sense-faculty and of the cloth of such a nature that several of the constitutive elements of the cloth are perceived, then, in this hypothesis, it would suffice to see the fringe of the cloth in order to see the whole cloth.
>
> If the Vaiśeṣika would say that in the case where one does not see the cloth when one sees the fringe, it is because in that case the central parts, etc., are not in touch with the sense-faculty, then this is to admit that one would never see the whole cloth, for its central-parts and end-parts which are supposed to make up the cloth are not perceived at the same time.

It's at this point in the passage that Vasubandhu points out the importance of time in this judgment. A central feature that make our object-based perceptions fail to match with the world is that objects are conceived in such a way that the element of time or change is abstracted away. This not only neglects changes to (and within) the object over time, but fails to match with the way that we actually perceive the world: as a series of sense impressions which take place over time. He goes on with the evocative example of a torch which is whirled in the air to create the impression of a circle of fire. It is only by ignoring the aspect of time that we can derive the image of a continuous ring of fire.

> If the Vaiśeṣika says that the central-parts and end-parts are perceived successively, then (1) this is to admit that the *whole or part-possessor* (the substance cloth) (avayavin) is not perceived;,(i) this is to admit that the conception of cloth or of straw mat has as its sole object the parts of the cloth cir of the straw mat, as this is obviously the case for the conception of a circle which we have of a circle formed by means of the circular movement of a torch in the air.

> Moreover: cloth cannot be anything other than the threads, for, in the hypothesis where it would be otherwise, when the threads differ in color, nature and arrangement, one could not attribute either color, nature, or arrangement to the cloth.
>
> If the Vaiśeṣika says that the cloth is varied in color, then this is to admit that that which is of different nature creates that which is of different nature; moreover, supposing that one of the sides is not multicolored, in looking at it one would not see the cloth [as it is] but see it rather as multicolored.
>
> But do you dare to say that the cloth, made of threads of varied arrangements, is of varied arrangements? It would be truly too varied to be one entity!

Having examined the individual elements which make up the cloth, Vasubandhu found those threads to have their own characteristics (requiring their own essence-bearing metaphysics, under Vaiśeṣika assumptions) but not to carry the essence or core of the cloth. If the cloth's own self-nature (svabhava) is to be discovered, it is not to be found there. In fact, many of the conceptual entities we assume exist in the world are highly variable over time, which casts some doubt on the accuracy of perceiving them as static objects.

> Moreover, consider this entity (avayavin), which is the light of fire (agniprabhā): since its heating and illuminating power varies from beginning to end, one would not be able to recognize either its color or its tangible qualities...
>
> For us, the atoms [paramāṇu], although suprasensible [atfndriya], become sensible [pratyakṣatva] when they are combined [samasta] in the same way, it is to the combined atoms that the Vaiśeṣikas attribute the power to create coarse bodies; in the same way, the factors of visual consciousness must be combined in order to produce a consciousness. (Vasubandhu et al., 2012c, p. 1111-2)

In this final section of the passage, Vasubandhu, is translated as referring to 'atoms.' This usage should be understood as referring to the philosophical definition of the word, not

as it is used in physics to describe matter only. In an early period of Buddhist metaphysics, respected by most surviving Buddhist sects, the goal of analysis is to reach the smallest level at which things cannot be further subdivided. In this sense, 'atoms' can also represent the smallest divisible bit of sense impression (such as perceived color, temperature, etc). In this way, it's the 'atoms' of sense impression that we interpret as a piece of cloth. In doing so, we are creating a work of mental fiction. Hopefully its a fiction that is conventionally useful in conducting our daily lives, but we shouldn't confuse it with objective truth.

Though it may be true that the self as a unitary whole is a conceptual fiction, that concept is only of broad theoretical interest unless specified further. In order to make this concept be of practical use to researchers working on artificial intelligence, it's necessary to have a basic working knowledge of some of the mental functions described. As Artemus Engle explains in their commentary on Vasubandhu's work:

> Classic Buddhist arguments assert that if the self were real, it would either have to be identical with some element of the five heaps or entirely distinct from them. By refuting both possibilities, we can conclude that the self is merely a nominal entity whose existence is dependently ascribed in relation to the five heaps. However, before we can comprehend these arguments properly, we must develop an adequate grasp of how the five heaps are meant to be understood. (Engle, 2009, p.14)

The heaps are organized as a careful deductive analysis of the different kinds of functions that human cognitions can perform, with an eye towards discovering the constituents of our congnitions down to the smallest possible units which are unable to be analyzed into smaller parts. These ultimate units to be derived were named dharmas.

As Vasubandhu argues, seeking an unchanging essence of a person is a fruitless task. No such self can be found among the aggregate mental entities that make up the function of a person's mind. Instead he says that a person is a useful conventional fiction which we use to designate an aggregation of mental functions, each of which is conditioned in its arising,

continuation, and destruction. This concept is often described in Buddhist philosophical literature as "co-dependent arising" and it applies to all mental functions, down the the most atomic unit of thought, the "dharmas" themselves. This "conditioning" means that observed mental functions must be explained through a causal story that includes interactions with other mental processes (cittas) and sense experiences from the world (rupa).

> "If the person is a real entity, it will be other than the aggregates, because its nature exists [then] on its own, since each of the aggregates is other than the others; [in that case], either i. it will be produced by causes [and then it will not be eternal as you say, and you will have to state its causes]; or else; ii. it will be unconditioned: and this is a non-Buddhist false doctrine; if it were unconditioned, the person is not able to "function" [or "be affected by anything or produce effects"]. It is thus fruitless to believe that the person is a real entity. 2. But if you admit that the person exists only on the level of a provisional designation, you abandon your doctrine and you side with our opinion." (Vasubandhu et al., 2012a, p.2526)

2.3 Process Metaphysics

This points to another aspect of the Buddhist description of mind which is useful to understand at the outset, and that is it's preference for process metaphysics. When trying to understand the basic framework for thought in a human mind, it's reasonable to ask the question "Is the world primarily made of objects or processes?" An object, in this sense, is a thing which is described by its form and its substance, and not by how it changes over time. A process, by contrast, is thing which is described by the conditions which allow it to arise, the conditions which allow it to persist, the internal changes that take place within the thing, and how the thing changes its environment. While the western philosophical tradition descended from Ancient Greece mostly focused on the former, Buddhist philosophy derides objects as mere conceptual fictions, useful, but not accurately reflecting reality.

The focus on object-oriented metaphysics is strong in the west, though not universal. Notably, the philosopher Alfred North Whitehead is the most prominent voice for this philosophy in the west. As Rescher summarizes, this branch of philosophy focuses on the illusory nature of objects, in that they fail to appreciate changes over time. "To be a substance, there must be an ongoingly self-identical bit of physical reality-a substantial core or essence-that assumes or discards properties over time. The paradigm of a core property bearer with variable properties is pivotal here. But just this sort of thing, a changeless core, is effectively impossible to come by in a world subject to pervasive and unremitting change in its substantial contents." (Rescher, 131) This notion of process metaphysics closely mirrors some ancient arguments in Buddhism.

The Buddha emphasized that one of the most important steps towards seeing the world as it is was to understand that many of the persons and things we encounter, we should realize that they are entities which are subject to constant change. As Ronkin explains

> The Buddha taught that to understand this repetitive condition in samsara is to see reality as it truly is – not a container of persons and things, but rather an assemblage of interlocking physical and mental processes that spring up and pass away subject to multifarious causes and conditions. In this respect it might be said that the Buddha had a distinctive, process epistemology: he taught that sentient experience is best understood in terms of dynamic processes that occur in a non-random order, and that to under-stand the causes and conditions of this dynamism is to gain insight into the way things truly are, which is equivalent to liberating knowledge.(Ronkin, 2005, p. 42)

So, part of what Buddhists mean by saying that there is no self is to say that there is no part of the self which remains unchanged through time. Instead, the self is marked by constant change, so any concept of "self" which does not reflect this change is in this sense "empty" of real objects. However, just because a concept is empty or illusory in this way doesn't mean that the concept is useless or should be dropped from our speech. "Empty,"

but useful concepts of static objects abound. This emptiness *(śūnyatā)* of an internally-determined nature *(svabhāva)* gives way to a kind of meaning determined by interdependence. Objects get their meaning through their process-based interactions with the rest of the world.

So, the argument that arises out of this Buddhist analysis of the world is that it is fundamentally built out of processes. Despite this reality of the world, we tend to understand the world in useful conceptual entities we think of as objects. These objects are generalized and idealized forms which resemble the processes, but are in some sense timeless, in the sense that their definitions don't cover the elements that refer to change over time. These conventional ways of thinking are useful, and I suspect that as we better understand algorithms that adequately mimic mental processes, we will find that understanding things as processes all the time is unreasonably computationally expensive. It may even be the case that systems of metaphysics oriented towards objects are associated with a different part of the brain and/or mind using different systems, such as the "fast" and "slow" systems argued for by Kahneman (2011). One common source of ambiguity within language may be that our language is more often oriented towards these useful conceptual entities, objects, but our meanings are disambiguated by reference to another mental system which has a process-based understanding. This process-based understanding may not be easily accessible by linguistic metacognition, perhaps even to the point of representing knowable, but ineffable truths. For the sake of brevity, however, I won't get too far into characterizing these ineffable truths in this work.

Although most of the things we usually identify as objects seem to maintain the same substance over time, it is not too difficult to think of cases in which a substance approach fails. My favorite example for the illustration of this concept is a standing whirlpool in a stream. Imagine a swiftly flowing stream which is interrupted by a rock which sticks out of the water. As it flows past the rock, the water moves in such a way as to produce a tornado-like swirl of water a foot downstream of the rock which will continue to exist as long as the local conditions in the stream (flow-rate, temperature, quantity of flotsam and/or ice)

remain steady. If identity must be based on substance, a whirlpool must be a constantly-changing substance, as it is made of water that is gone in a moment, only to be replaced with new water that follows the same pattern. While the substance seems to constantly change, the whirlpool still can be thought of as a thing, and worthy of a place in our understanding of streams and rivers.

A whirlpool such as the one I described should be able to be described with it's own identity criteria among the backdrop of other linguistic entities in natural language. We can certainly refer to a whirlpool as if it were a thing, point the individual pattern out to another person, and note its characteristics. If it's an object, it certainly can't be one by virtue of any part of substance that remains over time. A whirlpool is a pattern with a certain continuity over time though its substance is constantly changing. It doesn't even seem to necessarily have to have the same type of substance over time in order to qualify as the same thing: if the water filled with sediment the pattern could still continue, depending on how the pattern is understood or defined originally. If the stream, bizarrely, started flowing with vinegar instead of water (or some other substance with similar viscosity, weight, and anything else necessary to remain the same shape) the whirlpool could still remain a comprehensible linguistic entity throughout. The only thing that remains of the whirlpool through this process is its pattern, so if identity persists then this must be its source: a continuous pattern. Other obvious examples of things we could assign identities to which nonetheless can't be identified by substance are sounds and waves, which are identifiable despite only being patterns of movement in the underlying substrate of matter.

Vasubandhu gives an example of making a static object as a useful fiction based on the ever-changing nature of reality when he describes a circle of fire intuited from the impressions left by a spinning torch. We see the circle based on our repeated observations of the movement of the torch over time. He argues that we can observe ourselves creating these kinds of conceptual fictions when we form a conception of a mountain. Since a mountain is too large to be viewed all at once, our impressions of the mountain must be built up out of sense

impressions that we get at different times in our journey. "if it is thought that one sees the extended object-referent [mahatā] all at once [sakṛt], for example, a mountain [parvata], then that is by way of an illusion, because we see quickly [asuvrtti] the parts of the mountain: this is obviously so when we see the circle of fire drawn by a whirling torch [alatacakra]." (Vasubandhu et al., 2012b, p.297) So it is with our sense of self. We can observe ourselves using different mental functions at different times and from the aggregate of all of these experiences we create a conceptual fiction to represent ourselves. Vasubandhu is trying to point out that this fiction may not accurately reflect reality. Mistakes made in understanding the self can have wide-reaching consequences, as I'll argue in chapter 4.

Chapter 3: A Philosophy of Interdisciplinary Science

In this chapter I'll introduce the model of a scientific analogy that I'll be working with. I argue that this model of analogies can make plausible comparisons between domains, contrary to those who think that different domains of knowledge are incomparable. I'll examine the Kuhnean concept of incommensurability to show that incomparability is not a necessary result. I'll argue, contra Wallace, that analogies can have this effect without first having to settle questions like the ultimate nature of the substance of the mind, because a scientific analogy doesn't assume the truth of the compared domains and such considerations are radically underdetermined by current evidence. Finally, I aim to show that the the similarities we observe in analogical reasoning don't necessarily lead to unfalsifiable judgments, avoiding the worry of Popper. The model of analogy I will use follows Paul Bartha's approach to structuring analogies in *By Parallel Reasoning*. Bartha (2010) His approach to analogies discusses how to draw out the relevant similarities between domains, acknowledging any relevant ways that the domains are disanalogous, and has the result of creating a detailed account of similarity and dissimlarity which can be used to inform scientific judgments.

In writing on this subject, I assume that analogies are an important part of doing science in that they can help to compare disparate domains. Bartha, who devotes much of his book to the explanation of successful applications of analogical reasoning in science, notes some prominent examples here: "One famous example is Maxwell's discovery of the displacement current, around 1860, by modeling electromagnetic phenomena with a mechanical configuration of rotating vortices and idle wheels. In mathematics, too, striking breakthroughs have involved analogies between such diverse fields as logic and topology." (Bartha, 2010, p.2) Given that analogies can, by their very nature and definition, compare two disparate

concepts, analogies are a natural place to begin when discussing an interdisciplinary "contemplative science," like the one proposed by Alan Wallace Wallace and Hodel (2009a). Coming up with a brand-new and convincing scheme that plausibly describes analogical arguments is a grand and daunting task. Fortunately, I have no need to accomplish such a thing as worthies such as Paul Bartha and Mary Hesse have already made excellent strides along these lines. It would also be an amazing accomplishment to prove something definitive about the mind and how it is related to the brain from my armchair. Alas, that may be another task which is beyond me, though. Instead, I think I should keep my ambitions limited to those described by Hesse here:

> It does not, of course, follow that such an investigation will provide anything like an infallible method for the construction of theories, any more than it is the intention of accounts of methods of induction to provide infallible induction machines. All that is being attempted is an analysis of what assumptions are made when analogies are used in science, and how it is that certain hypotheses rather than others suggest themselves 'by analogy.' Whether the hypotheses thus suggested turn out to be true is, as always, a matter for empirical investigation. (Hesse, 1965, p.55-6)

In that vein, it bears mentioning at the outset several things that I will not be doing in this chapter. This chapter does not establish a deductive algorithm which somehow computes an analogy and yields a certain result. For reasons I'll go into in the chapter, a deductive strategy such as that is unlikely to yield effective results. I also won't be establishing an inductive scheme for analogies that yields a clean probability number for how "like" one concept is to another. Instead, this chapter grapples with the complexity of the analogical reasoning method, relying as it does on complex judgments of relevance and similarity. I'll try and do so in a way that serves three goals: The first is to clarify how scientifically relevant comparisons can reliably result from the process of making a careful analogical argument. The second goal is to clarify a thorny issue that artificial intelligence researchers encounter in

natural language processing (NLP). Finally, this chapter will lay the groundwork to discuss the logical problems that arise when definitions change in meaning across the two domains compared by an analogy. I will later argue that a subtle shift in definition between what a researcher in Cognitive Science might think of a term like "mindfulness" and what a Buddhist practitioner means by the word may prevent them from conducting fruitful interdisciplinary research.

My work on analogies is informed by work one in the Advancing Machine and Human Reasoning Lab (AMHR) at the University of South Florida. In the lab we created a model of analogical reasoning entitled "Warrant Game - Analogy" (WGA) to guide human users through the process of creating a good, structured analogy. What constitutes a good, structured analogy is the central issue I'm addressing in this chapter, and I'll argue that it is an essential component of creating an effective philosophy of interdisciplinary science, as well as essential to progress in AI.(Bartha, 2010, p. 64) In structuring the model of analogies we relied most heavily on Paul Bartha's explanation of the reasoning process in By Parallel Reasoning, so I'll begin with an explanation of the features of that model.

3.1 Plausibility and the Analogy Model

When you're creating an analogy, it can be difficult to discover if your analogy is a good one or not. It isn't the sort of judgment which we can easily label as 'true' or 'false' and it would be strange to argue that an analogy has a specific chance of being true, measured as a probabilistic judgement of likelihood of truth. Instead, as I'll argue in this chapter, the point of creating an analogy is to come up with a plausible generalization which is supported by isomorphisms between the subject and target domain.

Among other things, the plausibility of an analogy depends on the strength of the prior associations represented (e.g. How closely are the things inside each domain interrelated?), extent of the positive analogy (e.g. What do we know the two domains have in common?) and how it relates to other analogies (e.g. Is there a better fit for this dynamic elsewhere, in

a different source domain that would make a more plausible analogy?). Some considerations related to the plausibility of an analogy might be important enough to dismiss the argument, defeating it on its own. One consideration that is particularly influential is compatibility with other, related theories which are already well accepted. However, it's frequently the case that some judgment might make an analogy less plausible, but not make it entirely implausible. Paul Bartha argues that these factors show that analogies aren't always clearly true or false, but are defeasible on the basis of complex judgments.

> One of these factors—compatibility with accepted theories—can and should be incorporated into our assessment of an analogical argument because, as we saw in the Rutherford/Bohr example, incompatibility will defeat a conclusion about plausibility. The other considerations, however, act not as outright defeaters but as factors that influence overall degree of plausibility. There can be different models for assessing the overall degree of plausibility of a hypothesis. You might employ a weighting function that combines different components: analogical support, simplicity, and so forth. (Bartha, 2010, p. 6)

The defeasible nature of the argument is the result of evaluating theory on the basis of the desiderata of theory evaluation, as discussed by Kuhn and others. "In summary, when thinking about the role of analogies in science, the most significant issues in evaluating computational theories are predictiveness, applicability, scope, and simplicity. All are familiar from philosophical discussions of choice between scientific theories." (Bartha, 2010, p. 62) Just as with scientific theory evaluation more generally, an analogy must be judged, according to this view, by how well it fulfills several competing desiderata. These desiderata are competing in that there are often trade-offs between these different criteria. There is no uncontroversial, algorithmic way to apply these desiderata, let alone weigh them against each other.

Even when an analogy has prima facie plausibility, the significant pairs in the negative analogy might cause us to doubt it. After all, a better analogy (which is partly determined

by the weaknesses in negative analogy) might lead to an incompatible conclusion on the same topic. As Kuhn notes, a hesitation to do the complex mental calculations required may contribute to older researcher's reluctance to re-evaluate their theories. After all, understanding the complexities of modeling analogies doesn't come easily to everyone, and it's easier to remember one's past dismissals of alternative theories than it is to carefully read and re-evaluate what you've previously concluded. As Kuhn noted, sometimes a theory must wait out these established research scientists, in the hopes that the theories they refuse to re-examine may die out naturally. If we could give a definite (binary or numerical) procedure for evaluating theories, we could avoid this problem, but given the complex nature of the desiderata, this is a problem that will simply need to be accepted.

3.2 Advocate and Critic

Paul Bartha proposed that analogical arguments, once proposed, may be best developed by iterative improvements to the argument by two participants. This gives the strategy for analyzing arguments of this type a structure that is similar to a game, with the two sides fulfilling roles he describes as the advocate and the critic. These roles are both cooperative with one another, in that they are both working together to improve the quality of the analogical argument, but they are also in competition with one another. This competition stems from the fact that the two roles are trying to improve different aspects (the analogy's explicitness or its economy, in Bartha's terminology) of the analogy, and improving those disparate aspects often interfere with one another.

> ...we imagine that an enthusiastic advocate presents the analogical argument to a polite but moderately skeptical critic. Introducing this framework highlights the need to balance two competing pressures at work in representing and evaluating arguments from analogy: explicitness and economy. On the one hand, the critic wants the argument to be as explicit as possible, noting every factor that might be relevant to the conclusion, since the inclusion of detail increases the chance of

> exposing a weakness in the argument. On the other hand, the advocate wants to be economical about what counts as relevant. After all, a successful analogical argument shows that some differences don't matter. (Bartha, 2010, p.102)

These roles identified by Bartha are filled by players in our WGA program. Although the two roles are clearly at odds with one another, both participants are aimed at the creation of a successful analogical argument. We describe some of the potential perils from misunderstanding these roles in Assessing Evidence Relevance by Disallowing Assessment.

> ...these roles should be seen as collaborating in the creation of an analogical argument,even while they compete to determine the qualities of the resultant argument. Understanding these roles is important for avoiding the problems that might result from a straightforwardly antagonistic relationship, which treats the loss of the opponent as a goal to be achieved.A bad-faith advocate, imagining his duty to make a strong comparison, might refuse to focus in on an area of relevance. Instead, this advocate might try to draw a multitude of connections in the source and target domain, hoping to make the connection stronger that way. This would lead to an unhelpful list of similarities that cannot cohere to any rule. A bad-faith critic, in response, might refuse any and all additions to the source and target domains as irrelevant, at which point no progress could be made. These framing problems are arguably the result of the participants not appreciating the collaborative nature of the work. (Licato and Cooper, 2020, p.12)

It's important to note here that I'm not proposing that there is any uncontroversial algorithmic mechanism for determining explicitness or economy in an analogical argument. Indeed, the lack of such measures to apply is part of the reason we were forced to use human participants to iteratively improve an analogical argument, rather than having these roles fulfilled by software. Some of the currently insuperable difficulties involved in making these

determinations depend on judgments of similarity and relevance that rely on background information known to the speakers, as Bartha explains. "It is legitimate for the critic to require compatibility with widely shared background assumptions—presumably they are shared by both advocate and critic. Accepted scientific theory is part of this shared background, but other considerations, such as simplicity and competence, are not. They are interpreted differently and given different weights by different individuals." (Bartha, 2010, p.7) A successful analogical argument often draws on background knowledge that the two participants share. The argument is evaluated, however, on the basis of several complicated considerations which I'll introduce in this chapter as "desiderata." An awareness of these difficulties is not only necessary to properly evaluate analogical arguments, but will also serve to set up some of the problems of natural language processing (NLP) that will arise in later chapters.

3.3 Fact Pairs in the Source and Target Domains

The central purpose of an analogical argument is to compare two different, sometimes completely unrelated, situations or objects to look for homologies: the things they share in common. These commonalities can then be explained with a general rule which applies to both domains. In our work on WGA, we treat this generalization as a warrant as explained by Toulmin (2003b). The two participants accomplish their goal of developing this general warrant by articulating the prior associations in each domain which appear similar and assessing the relevance of each of these associations to the analogy. The source and target domains in WGA are represented by the left and right sides in figure 2.

Relations between fact pairs is largely the result of a judgment of similarity: The fact in the source domain mentioned must resemble its opposite in the target domain in some way. In Figure 1, the source is represented by the left-hand column, and the target is represented by the right. The source domain is understood to be a well-understood domain whose relationships will be applied to the target. The type of similarity focused on is not

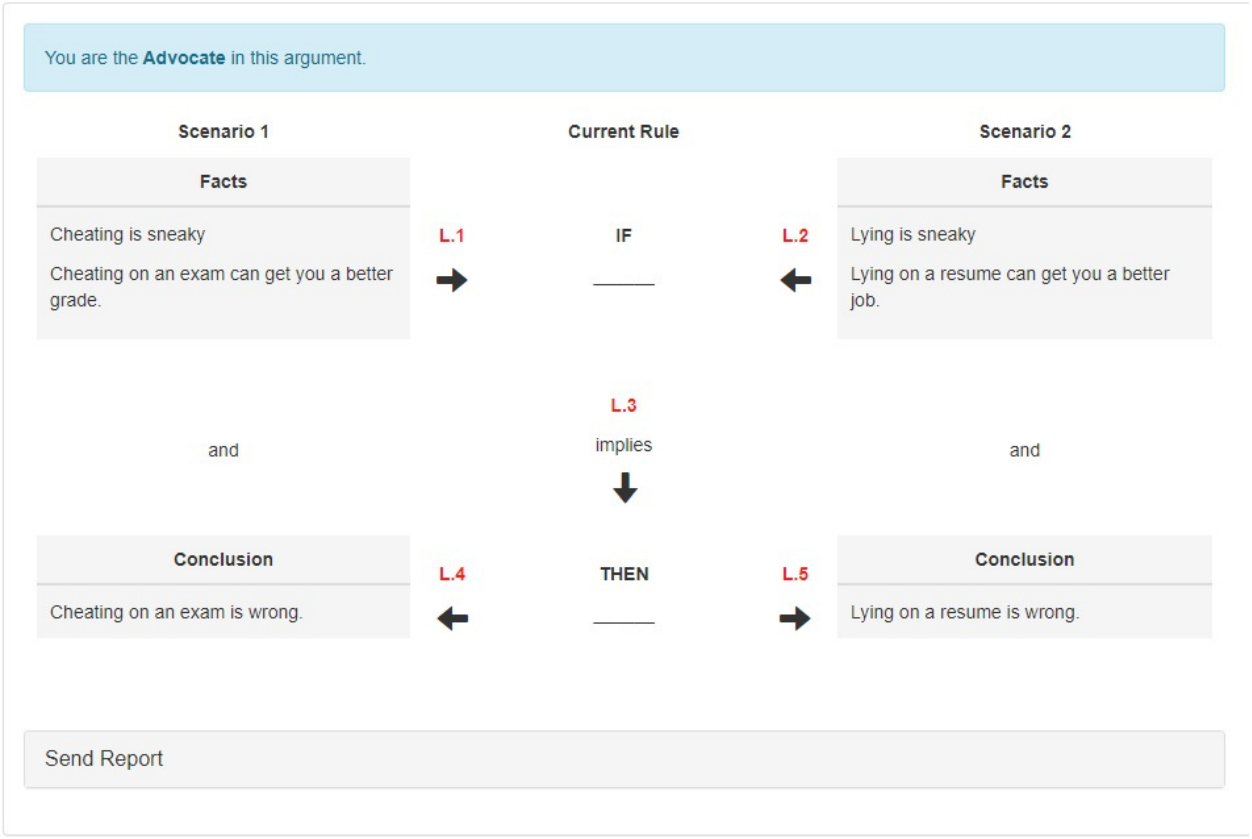

Figure 3.1: WG-A Layout

necessarily physical similarity. In fact, the two sides of the fact pair might be substantially different and may only appear similar after they have been idealized, restated, or generalized. The pair need not even resemble each other in themselves, but may only be similar because the two things play similar roles in their respective theories. For example, I might make an analogical argument that says the human eye is like the feeler appendage of an insect, despite the vast differences between the parts. The analogy works because an eye fulfills a similar function to the function served by the insect's feeler in at least one way: both are means of mapping the world around you. Similarity judgments are so flexible and sensitive to context and background information that computers struggle to make them. Making this kind of judgment requires human participants, given the current state of artificial intelligence research.

While similarity judgments rule assessments of how like one fact is to its paired opposite, relevance judgments compare the new fact pairs to previous fact pairs already in evidence. In order to be an improvement on an analogy, the new fact pair needs to be relevant to the situation at hand. After all, a fully complete assessment of just about any two objects were similar would find quite a large number of similar atoms, but using similar atoms in their construction is not relevant to many analogies. In order for a new fact pair in WGA to advance the analogical argument it must iteratively articulate more of the relevant prior associations that the argument relies on. The difficulty of assessing relevance is one of the central problems with inductive approaches to analogies, which I'll discuss in a later section. Bartha, in fact, names a number of different ways that a new fact pair might be relevant to an established analogy.

> The account developed in this book, by contrast, proposes a classification scheme based on the different types of vertical relations in the source and target domains. The vertical relations provide the clue to determining which similarities and differences are relevant. Different sorts of vertical relations naturally lead to different assessments of relevant similarity. (Bartha, 2010, p. 25)

This is a vertical relationship between new, added fact-pairs and previously accepted fact-pairs and is represented as such in WG-A. I'll detail the different types of relevance that Bartha identifies in the section on hypothetical analogies, as an understanding of the different ways that a fact-pair can be relevant informs the creation of interesting comparisons.

The model of analogy we use in WG-A borrows it's terms from Keynes, though they have been adopted by other researchers since Keynes (1921):

- Positive analogy (P) - Propositions Pin the source domain and P'in the target domain that correspond to "known similarities".
- Negative analogy (N) - Proposition groups A, $\sim$ B in the source domain and $\sim$A',B' in the target domain corresponding to "known differences" between

the domains. For example, the facts "Earth has an atmosphere" / "Mars does not have an atmosphere"would be in A and ∼A', respectively.

- Neutral analogy (O) - A set of propositions in the source such that the truth values of analogous propositions in the target are not known, and vice versa.
- Hypothetical analogy (Q) - A single proposition Q known to hold in the source and a hypothetical proposition Q'in the target whose truth value is not known, but is the conclusion of the analogical argument (Licato and Cooper, 2020, p. 2)

I'll explain how each of these elements contributes something essential to the proper representation of an analogical argument throughout this chapter.

3.4 The Character of Science and the Goals of Analogy

Before getting into the details of analogical structure, it's worth wondering first if the kind of interdisciplinary analogy comparing cognitive science theories with Buddhist philosophy is advisable to begin with. In this section I'll discuss scientific values, which will help us make the determination of whether an analogy is plausible, or which of two analogies is more plausible to see if there is anything that would prevent them from functioning in this case. Paleontologist Stephen Jay Gould, in his book Rocks of Ages: Science and Religion in the Fullness of Life famously defined the distinction between science and religion by declaring that science is concerned with epistemic reasoning, but that religion focused instead on values and purposes. He concluded that the two realms covered different subjects with different methods, and cast doubt on whether one would ever be relevant to the other. He wrote that science and religion operate in non-overlapping magisteria that nonetheless need to be integrated in some way to lead a full life. To make this case, he argued that science concerns empirical data, and religion, by contrast, concerns human purposes, meaning, and value. "...the net, or magisterium, of science covers the empirical realm: what is the

universe made of (fact) and why does it work this way (theory). The magisterium of religion extends over questions of ultimate meaning and moral value. These two magisteria do not overlap. . . " (Gould, 2007, p. 10) His argument is significantly complicated, therefore, by the fact that epistemic reasoning includes values. In addition to the value considerations that shape the ethics of scientific testing, epistemic values are involved as well. Kuhn identified accuracy, consistency, scope, simplicity, and fruitfulness as values to which scientists must be committed. (Kuhn, 1977, p. 322) These goals are shared values of the community of scientists which guide the actions of the discipline.

> Values like accuracy, consistency, and scope may prove ambiguous in application, both individually and collectively; they may, that is, be an insufficient basis for a shared algorithm of choice. But they do specify a great deal: what each scientist must consider in reaching a decision, what he may and may not consider relevant, and what he can legitimately be required to report as the basis for the choice he has made. . . Different creative disciplines are characterized, among other things, by different sets of shared values. (Kuhn, 1977, p. 331)

Kuhn argues that different scientists can place different emphasis on different values, so a different emphasis is not enough to make things unscientific, though different values might. In the context of this discussion, then, it's reasonable to worry if Buddhism imposes a different set of values which are inconsistent with the values of science. This is a serious concern for incommensurability, as I'll argue later, because the values might shape the characterization of problems and their accepted solutions to the point of coloring the definitions of key terms.

Do Buddhist theories of mind serve these epistemic scientific values? I believe that a case can be made that they do. One worry that someone might have in this conversation is that Buddhist theories of mind might hold faith-based views not just in the absence of evidence, but in the presence of contradictory evidence. This worry can be stated in terms of the above scientific values by saying that Buddhist theories of mind might not value consistency with new epistemic evidence. After all, if neurology produces evidence that seems to disprove

something in a Buddhist theory of mind, isn't it plausible that Buddhist practitioners will ignore the contrary evidence? Fortunately, at least one sect of Buddhism has made a strong endorsement of this kind of consistency. The 14th Dalai Lama, head of the Tibetan school of Buddhism has said that "...if scientific analysis were conclusively to demonstrate certain claims in Buddhism to be false, then we must accept the findings of science and abandon those claims." (Dalai Lama, 2010, p. 3) One thing to note about this assertion is that the 14th Dalai Lama is cautious to lay down strict criteria for falsifying Buddhist belief, and about wider spiritual claims it would be difficult to imagine what that falsification would look like. In particular, he points out that science failing to find something is not proof of that thing's non-existence. He has explicitly argued that Buddhism must be willing to abandon elements that conflict with scientific discoveries. (Dalai Lama, 2010, p. 80) His statement is also compatible with a key feature of science: that the scope of what science can prove is expanding. Though scientific encroachment on traditionally religious principles may give some readers pause, it's a necessary and unavoidable consequence of studying subjective internal states through cognitive science. Discovering how to fruitfully compare different fields as they approach similar problems is, therefore, an important task.

The accuracy characteristic that Kuhn points to suggests that scientists should hew closely to epistemic principles. The 14th Dalai Lama is careful to explain, in his book The Universe in a Single Atom, that Buddhism has a longstanding respect for epistemic principles. "...in science, in the final analysis, it is empirical evidence that represents the last court of justice. At least in principle, this is true also in Buddhist thought, where it is said that to defy the authority of empirical evidence is to disqualify oneself as someone worthy of critical engagement in a dialogue." (Dalai Lama, 2010, p. 76) Indeed, there is a hierarchy of established fallacies that places accuracy with regard to physical evidence as a higher value than consistency with established Buddhist theory: "There is a dictum in Buddhist philosophy that to uphold a tenet that contradicts reason is to undermine one's credibility; to contradict empirical evidence is a still greater fallacy." (Dalai Lama, 2010, p.

80) So, it seems there are good grounds for arguing that Buddhist theories of mind can share in the scientific value for accuracy.

The other two principles that Kuhn describes as essential to the scientific enterprise are fruitfulness and simplicity. It's easy to make the case that efforts are being made by Tibetan Buddhists to make their conversation with science fruitful. It would be hard to see what else could motivate the establishment of the Mind and Life Institute and the books they publish without a respect for the fruitfulness of the dialogue. What remains is a respect for simplicity within a scientific theory, an application of Occam's Razor, and a dictum to remain true to previous theory where possible. It is on this value that the challenge is most difficult to answer, since Buddhist theories of mind posit exotic causal forces emanating from exotic supernatural entities and forces. However, given a piecemeal approach, treating some parts of Buddhism as more consistent with science than others, these problems can be omitted from the conversation with scientists.

Even if my argument as to the compatibility of Buddhism with scientific values is plausible, there remains a worry that Buddhism applies some additional values to which science does not ascribe. These values may still color the puzzles and terms of Buddhist theories of mind to the extent that they are incommensurable with neurology, for example. The worry that Buddhism applies additional values is considerable, given that the central focus of Buddhism is it's soteriology: it promises freedom from suffering. "Buddhism is a distinctive normative theory, spiritual practice, and/or practical philosophy whose First Noble Truth, its very first insight into the nature of life, is that suffering is abundant. The First Noble Truth of *dukkha* is normally stated this way: Everything is unsatisfactory and is, or involves, suffering." (Flanagan, 2011, p. 20) This can be seen as an additional value which sets Buddhist study apart from science. Since Buddhist theories of mind are explicitly geared towards the relief of suffering, this value may come into conflict with the goal of accurately understanding or representing the functional elements of the mind.

However, relief of suffering is also a straightforwardly acceptable normative goal in medicine and psychology. This normativity is often presumed without argument, given that the goal of neurology and psychology is often to bring the mind to a "healthy" state.

> There is a principled reply that can work to deflate the objection: think of psychiatry and abnormal psychology texts, or of anatomy and physiology texts, or of surgical manuals. All these bleed normativity. Concepts of health, well-being, and proper functioning are required or assumed by these fields and they are normative, possibly taken-for-granted normative concepts, but normative concepts nonetheless. Is that an objection to these texts and the fields they represent? Even engineering is normative. The principles of structural engineering enable us to build bridges and skyscrapers that last. That is what structural engineering is for. The fact that engineering is normative is not an objection to its status as science.(Flanagan, 2011, p. 104)

If freedom from suffering is already an accepted goal in disciplines assumed to be scientific, like medicine, it would seem unreasonable to insist on a different, value-free standard for Buddhist theories of mind to participate in scientific discussions. Even in the realm of ultimate values, there seems to be no fundamental inconsistency with science, then.

3.5 Positive analogy

The most obvious and essential element in analogical reasoning is what the two areas being compared have in common. These elements of commonality are often left unspoken; as enthymemes in an argument. In producing WGA, we tried to encourage participants to iteratively add to and improve how these commonalities are represented in our positive analogy section. For example, atomic interactions (especially those without ionic charges) are often compared to billiard balls bouncing off of each other on a table. "Motion and impact, on the other hand, are just the properties of billiard balls that we do want to ascribe

to molecules in our model, and these we can call the positive analogy." (Hesse, 1965, p. 8) Though there are obvious differences between atomic interactions and billiard balls, such as scale, there are also similarities between the two. In order to make the analogy scientifically useful, a judgment needs to be made about which similarities between the two domains are relevant.

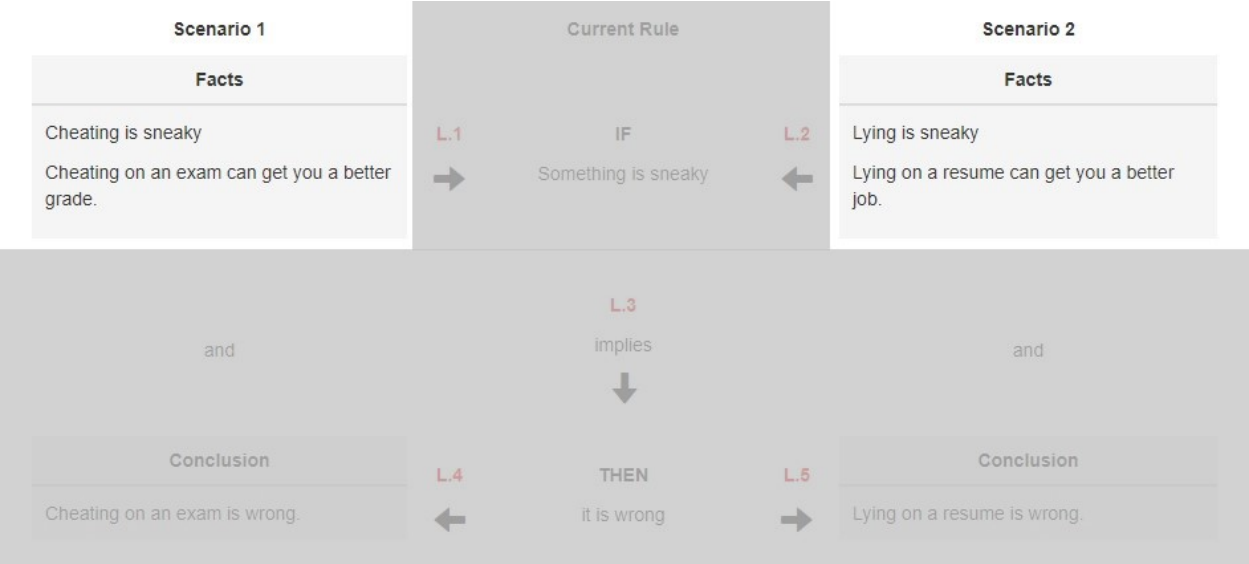

Figure 3.2: WG-A Fact Pairs

Thus, judgments of similarity and relevance are essential to fulfilling the basic function of analogies. In WGA these judgments are made by the human participants, due to the fact that they have not yet been made algorithmic enough to be accomplished by artificial intelligence. In this analysis I do not propose to solve these thorny issues, but an explanation of them will serve to set up problems in NLP to be addressed in later chapters: The demands of similarity and relevance, and how human reasoners carry out these tasks requires a detailed model of linguistic meanings and associations, a model which is currently beyond anything AI researchers have built.

In order for an addition in the source and target domains of the positive analogy to improve the analogy, the elements in the source and target domain must be similar. That must be true, but is still unacceptably vague. What kinds of things count as "similar"?

There are diverse ways that something might be considered similar to something else, and the concept may not even allow for a full definition. After all, even if we were to categorize every measure on which two objects or entities might be considered similar, a new measure of similarity may be defined tomorrow. Due to the open-ended nature of similarity, in WGA we left this judgment entirely up to the advocate and critic. In the course of evaluating an analogical argument, the two participants iteratively improve these fact pairs which may increase the similarity shown by matched facts. One common way to show that the two fact pairs are, in fact, similar is to apply the same predicate to both sides. If "S1" is a fact in the source domain, "T1" is a fact in the target domain, and P1 is a predicate that applies to both, then "S1 is P1" and "T1 is P1" would constitute just such a fact pair. Even if we regimented the process so that the only acceptable fact pairs are ones that share the same predicate, substantial ambiguity may still remain. This is the problem of "open-textured" predicates. Bartha explains that this problem may lead to reasonable disagreements about the quality of the analogical argument:

> A judgment of similarity is a judgment that two things are the same in some respect. It is commonly expressed by applying a single predicate to both things. Any such judgment can prompt a reasonable demand for justification if the application of the predicate is not a matter of definition or routine. In such a case, we are dealing with an open-textured predicate. Typically, the decision about whether the predicate applies involves a nontrivial comparison to paradigm cases or prototypes." (Bartha, 2010, p. 9)

These open-textured predicates pose particularly thorny problems when the logic of language is discussed in detail, particularly when it comes to restatements of natural language in formal logic and in the artificial intelligence specialty of natural language processing. Although clearly defined predicates can be understood as sets of objects which fall under the class that the predicate represents, it's less clear how we might deal with a predicate which is defined by paradigm cases.

Another key difficulty in judging the similarity of the source and target domains is distinguishing relevant similarities and dissimilarities from relevant ones. Relevant assertions of positive analogy will have a prior association that is causal or logical. In terms of the structure of WG-A, this consideration is in play when considering whether it is relevant to add a new fact-pair to the source and target domain. In order to be considered relevant, a new fact pair must relate to the situation at hand, either by being part of a single causal narrative or because one fact pair logically entails another. Hesse argues that models should be stripped to their essential properties, cutting out extraneous, irrelevant information.

> The important question clearly is, when is a property of a model essential? No clear-cut answer can be given to this, but various central considerations may be suggested. First, properties, which are causally closely related to the known positive analogy in the model are essential. For example, if the causal relations within the positive analogy are mechanical, mechanical properties are essential, but properties such as color, absolute size, etc. may not be. (Hesse, 1965, p. 90)

Here, Hesse explores the extreme theoretical difficulty of determining relevance to the analogy. In an analogy concerning the mechanics of a situation, you would introduce a new fact-pair by exploring the things that are causally connected to facts already accepted into the source and target domains. Not just any causal connection will do, however. The causal connections pointed to must relate to the overall generalization that will be drawn from the analogy, and it's not obvious that there is any uncontroversial way to decide how to frame the problem. Determining relevance is not a process that is well-modeled by computers, so that task is fulfilled by humans in our experimental program.

The role of eliminating non-essential propositions in an analogy is fulfilled by the critic in WG-A, who is motivated to reveal irrelevancies during the course of the critic's "attacks" on the statements already in place in the analogical structure. This is meant to contribute to the iterative improvement of the analogy because irrelevant additions to the formula will likely fall to subsequent attacks. WG-A represents the different domains being compared

in columns on the left and right side of the structure, so the inter-domain considerations of relevance are vertical relations between the facts.

These relevance considerations extend beyond the "positive analogy" section, though. The positive analogy is where fact-pairs which show similarity between the two domains are organized, but it's also important to note that a complete analogy would explore areas where the two domains are relevantly unlike one another as well. The term "negative analogy" is used to describe these relevant negative comparisons (though "disanalogy" is often used alternatively.) Though the simplified model of WG-A doesn't represent these negative factors, they are essential to a complete accounting of analogies, especially in situations where you're asked to judge which of two analogies is more plausible. A third type of relationship between facts in the source and target domain needs to be represented as well: in some areas we know the truth in the source domain, but not in the target domain. These kinds of facts are called either "neutral analogy" or "hypothetical analogy," and pairs there often suggest new and fruitful similarities between two domains which might affect the course of future study. These other sections of the analogy are also subject to considerations of relevance, as Hesse points out:

> A second consideration determining that a property is essential, then, is that it is causally so closely related to the rest of the neutral analogy, that the whole of this would become part of the negative analogy if the property in question were shown to be so. Third, it may be suggested that so long as some neutral analogy remains unaffected, the model may be retained in spite of encroachment of the negative analogy into essential properties, but the license allowed to a model in this respect will depend all the availability of alternative models.(Hesse, 1965, p. 91)

Here Hesse argues that we should only take into consideration similarities, dissimilarities, and hypothetical similarities that are relevant to the current situation. After all, there is usually an endless list of ways that any two things are dissimilar to one another, such as

the number and exact configuration of the atoms which make up the object. Further, she points out that additions to the negative analogy don't necessarily disqualify an analogy in the absence of a better comparison.

3.6 Negative Analogy

This brings me to an aspect of analogies that is often overlooked: the negative analogy. The negative analogy is a collection of ways in which the target and source domains are dissimilar to one another. When we make an analogy, claiming that something is "like" something else, we're not claiming that the source and target domains are identical. In order to make this comparison fruitfully, we need to be aware not only of the ways that the two domains are relevantly similar, but also keep in mind ways that they differ. These points of divergence help to limit the scope of the generalization that the analogy warrants and speak strongly to the overall plausibility of the analogical argument. As Hesse points out, this is particularly clear when examining scientific analogies, such as the model of billiard balls to approximate molecular collisions.

> When we take a collection of billiard balls in random motion as a model for a gas, we are not asserting that billiard balls are in all respects like gas particles, for billiard balls are red or, white, and hard and shiny, and we are not intending to suggest that gas molecules have these properties. We are in fact saying that gas molecules are analogous to billiard balls, and the relation of analogy means that there are some properties of billiard balls which are not found in molecules. Let us call those properties we know belong to billiard balls and not to molecules the negative analogy of the model. (Hesse, 1965, p. 8)

To make the comparison between billiard balls and molecules is not to assert that the two objects are precisely alike. If that were the case, then after a successful analogy we would consider a molecule a kind of billiard ball or vice-versa. Saying that the two domains

are analogous is to argue that they are similar in some ways, captured by the generalization encapsulated in the warrant, but it is also implicitly to say that there are many ways in which the two domains differ. As a result, a model of analogy needs a negative analogy section to be explicit and complete. As Hesse notes, this element is vital to using analogy in a way that yields useful scientific results. "...in seeing how this theory can yield new predictions in the domain of the explicandum it is often necessary to take account of differences as well as similarities." (Hesse, 1965, p. 98-9) In order for the billiard ball model above to adequately model the physical situation it's important, though in this case trivial, to understand that the hardness and shininess of the billiard balls doesn't indicate that molecules are similarly hard and shiny.

Something to note about the negative analogy is that it also suffers from the central difficulty of determining what kinds of dissimilarities are relevant to mention in an analogy. In the above example, it's the dynamic movements of molecules relative to one another that is similar to the interactions of billiard balls. The hardness of the ball is a characteristic which is relevant to the movements of balls on the table, but not one that clearly applies to molecules. Though molecules can bounce off one another in the manner suggested by the dynamics of billiard balls, some molecules can form chemical bonds with one another. There is no clear parallel to that kind of interaction among billiard balls. Another dissimilarity between molecules and billiard balls, one too obvious to be mentioned in most contexts, is that the two types of objects have very different sizes.

This poses a challenge to models of analogy like WG-A which try to articulate the necessary background knowledge to carry out an analogical argument; it's difficult to know when you have enough of the background knowledge represented to stop articulating. Negative analogy is also highly important to making determinations between analogies, since relevant ways in which the two domains are dissimilar tend to harm the overall plausibility of the generalized similarity.

3.7 Alternative Constructions of Analogy

There are two prominent approaches to interpreting analogical arguments as defeasible arguments: deductive models of analogy and inductive ones. In this section, I'll discuss both of these groups of theories and argue that either method is plagued by insuperable problems. In Mary Hesse's classic work, Models and Analogies in Science (1966), she argues for a deductive model of analogies. She argues that it would be possible to formalize analogical reasoning by defining key terms and statements and and explaining the relationships between the established terms. She argues that once these two factors have been established clearly, a successful analogy model would be able to algorithmically designate whether the analogy is successful or not.

> It may be that when the nature of the similarity is pressed, it will be admitted that the analogues do not both have the identical property B, but two similar properties, say B and B', in which case the analysis of the similarity of B and B' repeats the same pattern. But if we suppose that at some point this analysis stops, with the open or tacit assumption that further consideration of difference between otherwise identical properties can be ignored, we have an analysis of similarity into relations of identity and difference. (Hesse, 1965, p.70-1)

Her controversial premise there is to suppose that the operation of the analogy can be captured in the definitions and relationships of the formal system. As a result, she is committed to the view that once these elements have been established, it is no longer necessary to keep other background knowledge in mind. In terms of WG-A, Hesse's properties B and B' represent a fact-pair, with B in the source domain and B' as the fact it is paired with in the target domain.

Paul Bartha objects to these features of Hesse's (1966) model. He argues that there isn't a point in the analogical argument at which we can safely idealize some of the similarities and ignore others. He raises two principal objections to this approach. The first is that it

would be very difficult to decide when it would be appropriate to say that all of the relevant similarities have been well captured by the formal system. Absent such a judgment, it's inappropriate to ignore further similarities and differences. The second objection concerns the difficulty of determining which kinds of similarity are important to the current situation.

> I shall mention just two difficulties with Hesse's argument. The first is that it is overly sanguine to suppose that the stopping point represents a situation where further differences can fairly "be ignored." In practice, the advocate for an analogical argument often deliberately stops—indeed, often must stop—with identities that suppress important differences.... The second difficulty is that, contrary to Hesse's key premise, alternative analyses of similarity may be available." (Bartha, 2010, p.41)

The fundamental problem is that the relationships between parts of this difficulty becomes all the more apparent if it is true that analogies are to be evaluated like scientific theories; judged with a complex set of desiderata. In declaring that further similarities and differences can be ignored, the model loses the richness of the data which it may need to rely upon to judge the analogy in terms of desiderata. He admits, however, that this formalization wouldn't be a problem for every analogy, but that it causes problems often enough that this formalization shouldn't be taken as part of the generic model of analogical reasoning.

Macagno, Walton, and Tindale seem to agree with this point when they specifically dispute the assertion that there is a point at which further similarities and differences can be ignored through this formalization of terms and relationships. They draw on the complexity of material reality to support the idea that there is complex, relevant information in the makeup of a material object which shouldn't be idealized away. They make this emphasis on the holistic, relational nature of meanings in natural language while coincidentally drawing on my central example, of a red pen.

> "Inferences linking statements such as 'This pen is red; therefore it is colored' cannot be considered as purely logical, in the sense of being purely formalized according to the semantic system used in modern formal logic. Such inferences can be analyzed from a material point of view, relying on relations between the terms that are more complex than the number, arrangement of the terms, and the syntactic connectors . (Macagno et al., 2017, p. 224)

Here they suggest that whatever formalization we might use of an analogy, we are likely to lose information that is important to the operation of some analogies. Given that fact, it would be misleading to suppose that reasoning by analogy depends on these kinds of formalizations.

Another objection to deductive approaches to analogy is that our formalization of analogical arguments frequently include decisions made on the basis of human purposes, rather than objective facts about the world. In determining and articulating the prior associations of the facts in the positive, negative, and hypothetical analogies we're likely to categorize things and make determinations of relevance on the basis of our human goals. These considerations are in play in determining, for example, what counts as an "essential" property of a term, which must be expressed, and what is instead an "accidental" property which can be safely ignored. This is especially problematic for scientific arguments about analogy, as the analogical argument is supposed to show something about the outside world. Any threat to the objectivity of the information included in the analogy makes the efficacy of the argument suspect.

Deductive approaches tend to draw on this essential/accidental distinction, but the observations of pragmatism need to be kept in mind. How we define essential characteristics relates as much to our purposes as it does to realities in the world. Macagno, Walton, and Tindale make this point by describing the genera that define specific terms as pragmatic and dependent on the functions we find desirable.

> This generic "concept" is abstracted based on the specific relationship between the two terms of the comparison, namely the viewpoint that constitutes the purpose thereof. It does not correspond to the "essential" (or absolute, context-free) meaning (definition), but rather to a pragmatic, functional genus, setting out what the target and the analogous are for in the specific context. In this respect, essential and accidental similarities can be thought of as characterized by a similar process of abstraction.(Macagno et al., 2017, 230)

In other words, the distinction that we often make between essential and accidental properties does not hold for definitions within analogies, at least, if the analogies are going to be faithful representations of the world. In order for the genera represented in the analogy's fact-pairs to represent the world well it's necessary to include all of the information, both essential and accidental which affect how a physical object interacts with the world. Deductive approaches like Hesse's would leave off a lot of that important information. Though it would be nice to have a neat, algorithmic way to approach evaluating arguments by analogy, it seems that the deductive approach necessarily limits the amount of information available through its methods of formalization.

Another prominent group of theories treats argument by analogy as a kind of induction. This approach has some intuitive plausibility to it, since analogical arguments don't produce a clear "true" or "false" as their answer. Alternatively, viewing analogies as a kind of induction would suggest that the analogy produces a probability of being true, if only we can reduce it's parts to specifically weighted elements which can be related to one another. This approach, if successful, would mitigate our need to make some difficult similarity judgments, or at least allow us to shed our judgments of similarity once things have been reduced to numerical weights.

> The most important thing to notice about this way of representing the logical form of argument from analogy is that it makes no reference to the notion of similarity. The textbook accounts make argument from analogy seem highly

> objective. It looks like it represents a type of argument that can be evaluated in a scientific and objective manner using inductive reasoning to count up the properties shared by a set of entities to provide positive evidence supporting the argument from analogy and subtract the negative evidence of entities that fail to share common properties. There is no need for students to ask embarrassing questions about similarity. (Walton, 2014, p.7)

Though this approach is tempting, it suffers from two serious worries which I'll address here. The first problem is that there is no uncontroversial way to enumerate positive evidence and negative evidence. This naturally leads to the worry that irrelevant similarities or differences may add additional weight to one side or the other. The second is that inductive accounts of analogy tend to resemble induction from a single case, which is a notoriously unreliable form of reasoning.

Despite these problems, which I'll address in more detail below, there are also reasons to favor an inductive treatment of this problem. The inductive approach promises to be able to determine the weight of different aspects of the analogy in order to make the comparison between sides and between analogies proceed smoothly. According to inductive models of analogy, if it were possible to assign fixed weights to different aspects of the problem, it would be possible to discard the complexity of the background information thereafter. Walton points out the practicality of such a system, if successful, both in judging an analogy and comparing it to a rival analogy.

> If we could use numbers of this sort to calculate the strength of an argument from analogy, the argument could rightly be classified as inductive, as they advocate... To comparatively weigh up the strength of the one argument as compared to the strength of the opposed argument, we have to bring in factors that identify the respects in which one case is similar to the other, and have some device for estimating how similar one is to the other by attaching weights to similarity. But there is always the problem of how misleading it might be to attach numbers to

> the weight of importance each factor should have in a given case. (Walton, 2014, p.8)

Though much of the appeal of the inductive approach lies in its ability to avoid difficult judgments of similarity, Walton points out that these same judgments are the ones we'd need to determine the weights in the inductive approach anyway. In this way, inductive models of analogy don't actually avoid any problems of similarity, so much as they move them to a specific part of the problem (the assignment of weights) and ignore them afterwards.

Even so, there might be some advantage to corralling all of our similarity judgments into one part of the problem, even if those judgments are not thereby reduced in complexity. After all, if it's possible to weigh analogies against each other by their numerical weighted scores, that would be much simpler than taking into account all of the complexities of the background knowledge. This is particularly appealing from an AI perspective, as the direct comparison of numerical scores is a task that computers perform with excellence and precision. One problem with this approach is that it's hard to imagine any scheme which could uncontroversially decide on the relative strength of the weights applied. As Bartha points out, our schematization would frequently be related to our all-too-human goals and interests, rather than being oriented towards truths in nature.

> The problem, of course, is to determine when an unstructured list of matching features supports an inference to a common kind. There is no good general philosophical solution to this problem. Quine (1969, 1973) suggests that natural selection has equipped us with an instinct for making good similarity judgments and singling out interesting kinds. There is certainly some promise for a naturalistic approach to this type of analogical reasoning, but there are also grounds for skepticism... we make many errors about the significance of properties and kinds. (Bartha, 2010, p.199)

This pragmatic worry casts doubt on the effectiveness of classifying scientific analogies inductively. However, even if it were the case that we could uncontroversially give weight to different elements of an analogy, it does not appear to be true that there is any point at which we can discard background knowledge in favor of these weights. If it's right that analogies have to be judged and weighed against each other by something like the desiderata which determine scientific theory choice, then it's likely that there would be complex trade-offs in determining which analogy is more plausible. What if one theory is more consistent with prior theories (using a desiderata of simplicity or conservatism) while the other is potentially more fruitful? If all desiderata have been monistically combined into a single numerical weight and we're not referring back to prior associations anymore, this kind of judgement is entirely impossible. Even if we assigned individual scores to every desiderata, it's not clear that a numerical score contains enough information to judge whether a particular trade-off is worth it. There are ongoing debates about how these desiderata are to be enumerated, but it's interesting to note that Quine, Kuhn, Carnap, and E.O. Wilson have very similar lists of desiderata for scientific theory comparison. These tools of theory comparison have also been proposed as the ideal criteria for determining the plausibility of analogical arguments as well. As Bartha writes, "...analogical arguments that satisfy the general principles of the articulation model strike an excellent balance between conservative epistemic values (such as simplicity and coherence with existing theory) and progressive epistemic values (such as fruitfulness and theoretical unification)." (Bartha, 2010, p.239) It's important to note that striking this balance between competing desiderata is a complex intellectual task which must be performed both during the construction of an analogy and any time the analogy is challenged by a rival analogical explanation.

Agassi (1964) makes the case for analogy as a method for generating inductive generalizations from an analogical pair. In his way of approaching the problem, an analogy invites us to apply a previously-held generalization to a new case (or pair of cases). Bartha notes this assumption as something that doesn't apply to all analogies, and argues that this constitutes

a weakness in Agassi's approach. Essentially, Bartha argues that single case induction isn't the correct model of analogy because oftentimes analogies define a natural kind, not just confirm it.

> Based on the foregoing, the objection to understanding all analogical arguments as single-case induction should be obvious: it is simply too restrictive. Some analogical arguments may fit this pattern, but not most. In fact, successful analogical arguments may lead to the discovery of a natural kind... But we don't know that we are dealing with a natural kind when we make the original analogical argument.(Bartha, 2010, p.50)

If argument by analogy is just a case of induction on a single case, then it's fundamentally about drawing a relevant generalization of the single case in context, essentially using the generalization as an argument that the single case corresponds to the general kind identified in the warrant. This warrant, though, doesn't need to fit a previously established generalization and may be used to establish a new general kind. "This reflects a general truth about analogical arguments,: they may reflect the application of an underlying rule or generalization to two separate cases, but they do not have to presuppose such a rule." (Bartha, 2010, p.273) The warrant that an analogical argument justifies isn't always a pre-existing category to be verified through inductive reasoning, but often a category that is actively defined through the use of the analogy.

I take these two problems of inductive models of analogy to be insuperable. The appeal of an inductive approach is to, in some way, reduce the complexity of the problem by establishing numerical weights to different elements to eliminate the need for constant complex judgments of similarity and relevance, but this promise cannot be fulfilled if we judge analogies by desiderata which require the eliminated elements. Mary Hesse points out that even if we can't overcome these difficulties, we can still use these similarities to understand the overall plausibility of the analogy.

> Of course the description of similarities and differences between two analogues is a notoriously inaccurate, incomplete, and inconclusive procedure. Although we often feel some confidence in asserting the existence of a similarity and that some things are more similar to each other than to other things, we cannot usually locate discrete characteristics in one object which are positively and finally identifiable with or differentiable from those in another object. But the inconclusive nature of the procedure is not fatal here, because we are not looking for incorrigible inductive methods...(Hesse, 1965, p.76)

Though the simplicity of a true or false deductive answer, or a numerical probability generated by an inductive method would be preferable in many ways to a defeasible answer, it's not clear that any simpler methods would be effective in producing analogies which are relevant to scientific pursuits.

3.8 Incommensurability in a Model of Analogy

As I pointed out earlier, Masterman (1970) showed that analogies are consistent with some of the definitions of paradigm Kuhn used, and this was confirmed by Kuhn (1977). In this section, I'll explore Kuhn's view of analogies more fully and examine the application of incommensurability to my model of analogies. As Bartha argues, analogical reasoning is a key mechanism in the advancement of normal science, not only in scientific revolutions.

> As Kuhn has pointed out, even though the list of epistemic values important to science may be stable and widely shared, the precise interpretation and weight given to each such value varies from one scientist to another. We expect controversies about method, and we find them. Despite such variation, however, Kuhn identifies analogical reasoning as the main engine of expansion for normal science. Commitment to the core principles and problem-solving methods of a paradigm generates the confidence and creative force that leads scientists to extend those

> methods to new applications by means of analogies. In fact, the breakdown of analogical reasoning coincides with the point where normal science leaves off and scientific revolutions take place. But in normal science, analogical reasoning is the, or at least a, principal problem-solving method used by scientists.(Bartha, 2010, 246)

Since Kuhn argued that science relies on analogical reasoning so thoroughly, it seems clear that he didn't imagine that incommensurability made it impossible to make fruitful comparisons between domains. To use incommensurability to cast doubt on analogical reasoning would seem to be the kind of impermissible interpretation of incommensurability as incommunicability that Kuhn took pains to refute. To take this objection as a proof that different disciplines can't communicate is to ignore the efforts Kuhn made late in his career to address this problem and show that it's not fatal to his system of thought as represented in the Structure of Scientific Revolutions. Kuhn and Wilson (2001)

As I argue above, one of the central difficulties of incommensurability is the idea that the two paradigms under comparison might use the same word in subtly different ways. Leaving aside other issues concerning "paradigms," there is a clear way in which this criticism might apply to my analogical model generally. What if an analogy were to use the same word in both sides of a fact-pair, but the meaning of the word was different in each case? In such a circumstance, the similarity which makes them "paired" might be a superficial byproduct of our definitions, rather than representing a similarity that is grounded on similar features of the world.

Though I'm far from the first to note that the problem of differing definitions is a worry for analogical reasoning, I believe that most of the authors who have examined this issue have underestimated its complexity. Though it's possible to construct some basic analogies only using uncontroversially-defined words, making sure that each definition has necessary and sufficient conditions for inclusion under the term listed, this is not the only way that people communicate. Bartha notes some of the difficulties with word meanings that might

compromise analogies, despite trying to focus on "basic similarities" as driving analogical arguments. Here, Bartha draws on Kuhn's use of "paradigm"s in their analogical sense, by making reference to paradigm examples to fix the meaning of terms.

> Specifically, the basic similarities—those upon which the analogical argument is founded—are not treated as open-textured in the context of that argument. These basic similarities may be established by formal definitions or by a widely accepted classification scheme... To be sure, there are open-textured predicates in science. Many people believe that the extension of kind terms is determined (at least in part) by reference to paradigm examples (Kripke 1972, Putnam 1975, Kuhn 1979). Hesse (1966) argues that there can be theoretical concepts whose very meaning appears to be determined by analogy or metaphor. " (Bartha, 2010, p.9)

An open-textured predicate is a serious problem in this context, and ties in well with a discussion of Wittgensteinean family resemblances. An open-textured predicate has the ability to change the scope of its definition as a response to context clues. If I asked you to bring me the "red thing" from the next room, and you entered the room to find an American flag and a white flag, it would be natural for you to bring the American flag. Though the flag is only approximately 1/3rd red, it's the most red thing in the relevant context. On the other hand, if you entered the room to find a flag with a solid red color and an American flag, you would likely bring back the entirely red flag. In this example, the definition of the category "red" arguably shifts based on the available objects in the domain of inquiry, applying to the thing which is most similar to our concept of redness. Similarly, when a definition is based on a Wittgensteinean family resemblance, the extension of the terms is made on the basis of similarity to paradigm examples, and to the body of other words included under the term. From this perspective, changing any member of the set whatsoever changes the definition of the word.

The combination of these factors, and the potential to use words defined in these ways in normal analogical arguments, means that the worry of equivocation in an analogical argument is a serious one in natural language. The task of NLP is explicitly to engage with the complexity of natural language, not just artificial languages in which we can enjoy simple, uncontroversial definitions which are laid out algorithmically. Instead, it must try and make sense of the odd and often complex ways that people actually use language. A model of analogy which can satisfy these stringent requirements will have to deal gracefully with the significantly fuzzier reasoning we use when applying open-textured predicates which rely on Wittgenstenean family resemblance.

With these considerations in mind, it's possible to see the real worry of incommensurability as applied to an analogy. If a word is used (at least) twice in an argument, with a different operative definition in mind each time, then incommensurability appears to be a variety of equivocation in an analogy. The danger is clear, and further emphasized by the difficulty of participants in an argument to perceive this problem, even when well-trained.

> Because of their intrinsic lexical ambiguity, metaphors are extremely likely to cause the fallacy of equivocation and thus to deceive in the evaluation of the argument strength, i.e. the proper attribution of a certain analogy as its conclusion. In this sense, they might be particularly persuasive. Previous experiments suggested that participants have some difficulties in detecting a lexical ambiguity fallacy, especially when arguments are based on conventional metaphors and even when participants are experts, i.e. trained in logic and argumentation. (Ervas and Ojha, 2019, p.332)

So, overcoming worries about incommensurability in analogy are a serious concern, and it is one that will shape my argument about cross-disciplinary scholarship between Buddhist philosophy and Cognitive Science disciplines. In order to avoid problems of incommensurability, it's essential to pay careful attention to any word that appears in both the source and target domain to maintain consistency, to the extent possible, in the definition of the word.

This is one of the most important problems which I intend to address in the remainder of this dissertation.

3.9 Popper's Falsifiability and the 14th Dalai Lama

Another significant worry for interdisciplinary work between Buddhist philosophy and disciplines which participate in Cognitive Science, and that is that Buddhist philosophy can't be made scientific because it's pronouncements are unfalsifiable. This worry is enlivened by the historical record, since many of the people attempting the kind of interdisciplinary work are associated with the Mind and Life Institute, which is affiliated with Tibetan Buddhism. In The Universe in a Single Atom, the 14th Dalai Lama describes a personal friendship with Karl Popper, who proposed that falsifiability in it's predictions are what makes something scientific.

> During my first visit to Europe, in 1973, I had the honor to encounter another of the twentieth century's great minds, the philosopher Sir Karl Popper.... we struck up a friendship, and I. saw him again whenever I came to England, including a memorable visit in 1987 for tea at his house at Kenley in Surrey. I have a particular love of flowers and gardening, especially of orchids, and Sir Karl took great pride in giving me a tour of his own lovely garden and green-house. By this time I had discovered how great Popper's influence was in the philosophy of science, and especially on the question of scientific method. (Dalai Lama, 2010, p.33-4)

He recounts this long-standing friendship and it's clear that their discussions included robust explanations of this falsifiability thesis. This friendship and understanding of the falsifiability thesis was an important part of the founding of the institute, so it's clear that this concern is being taken seriously. This deep knowledge of the thesis is apparent when the

Dalai Lama notes some differences between this thesis and another epistemological principle of Buddhist philosophy:

> Popper's falsifiability thesis resonates with a major methodological principle in my own Tibetan Buddhist philosophical tradition. We might call this the "principle of the scope of negation." This principle states that there is a fundamental difference between that which is "not found" and that which is "found not to exist." If I look for something and fail to find it, this does not mean that the thing I am seeking does not exist. Not seeing a thing is not the same as seeing its non-existence. In order for there to be a coincidence between not seeing a thing and seeing its non-existence, the method of searching and the phenomenon being sought must be commensurate.(Dalai Lama, 2010, p.35)

In this quote he takes pains to point out that the falsifiability thesis is limited by the adequacy of the empirical tests the falsification is using. It's an important caveat to keep in mind when addressing Buddhist thought, which contains a lot of assertions which are not obviously falsifiable. Keeping this caveat in mind guards against a shallow scientism which would declare things that are not currently scientifically provable to be false.

Kuhn, of course, has differing views about falsificationism, but the desideratum of accuracy fulfills a similar purpose in his theory. In either case, the assertion is that science will make accurate predictions about the world, not just explain what has already taken place. In Bartha's own list of desiderata, he discusses this value in terms of "predictiveness."

> Predictiveness is concerned with whether the evaluation criteria embodied in both the program and the conventions about representation can be made explicit and yield a clear verdict about plausibility. The less clear the conventions about representation, and the more sensitive the program is to the details of representation, the weaker a theory's claim to be predictive. (Bartha, 2010, 61-2)

Predictiveness is the most important of the desiderata in creating an overall judgment about whether a scientific analogy is plausible. As a result of this, it is centrally important that any analogies used to compare Buddhist philosophy and Cognitive Science concepts be made with an eye to making the hypothetical analogy unambiguous in terms and applications, so that it makes a precise prediction about what will be found in the target domain.

In analogical argumentation participants rely on underlying isomorphisms between the two domains under consideration. This means that the problem is incredibly sensitive to changes in word-representation. As Hesse argues,

> ...there must already be a fairly well-developed system of relations in the observation language. The less developed this is, the more difficult it will be to ensure that an apparent isomorphism is not accidental or arbitrary. This means that the program will not be universally applicable and not applicable at all to observation predicates not already part of such a system in the observation language.(Hesse, 1965, 45-6)

In order to effectively run an analogical argument, it's essential to have a rich representation of word meanings and the relationships between words. People have this as part of our linguistic knowledge, but in an NLP application, an approximation of this representation scheme must be created from scratch. It is for this reason that I choose to examine some of the most difficult problems of word representation available in the literature, in order to find a scheme of representing word meaning which is powerful enough to operate in an approximation of human knowledge. In particular, I think that a successful word representation scheme for NLP would be able to handle an open-textured predicate which is defined by Wittgensteinean family resemblance. Since such fuzzy constructions appear to be in common use, being able to represent their structure is essential to representing the isomorphisms present in natural language.

Conclusion of the Chapter Analogical reasoning is commonplace, but often overlooked in the course of philosophical explication. It plays a key role in the creation and challenging of scientific theories, playing key roles in complex concepts like "paradigm," and has the potential to drive interdisciplinary comparisons. In this chapter I've attempted to explain a rigorous, clear way to analyze this kind of argument, at least in its simplest form. Along the way, I've described analogy as a defeasible form of reasoning which, like scientific theory evaluation, has to be conducted by balancing several different desiderata. Along the way, I've also briefly examined the prominent alternatives to constructing analogies in this way, and highlighted the problems which prevent these alternatives from being viable options. In doing so, I'm not arguing that every attempt at a crossover between Buddhist philosophy and cognitive science will be successful. Far from it, this analysis highlights the difficult standards that an analogical argument would have to meet to be considered scientific. I will keep these strictures in mind as I make my case that such a crossover, if carefully managed along the lines suggested by criticisms in the philosophy of science, can be fruitful.

In the following chapters I will make my case that Buddhist Philosophy provides a fruitful source domain for scientific analogies when a subject in Cognitive Science is used as a target domain. It should provide a structured, and relatively uncontroversial way for the two domains to interact without implying that either needs to uncritically accept the findings of the other. Analogies don't necessarily assume the truth of the domains that they are comparing, but instead rely on similarities in the causal and logical relationships that operate in those domains. In doing so, they are very sensitive to the ways in which words can be represented. I'll explore those issues of word representation from the perspective of later Wittgenstein, to explicate some of the thorniest issues that need to be dealt with, and discuss insights from Buddhism which elucidate the problem to some degree.

Chapter 4: Analogies Comparing AI and Buddhism

4.1 Interdisciplinary Success

Wallace, Kabat-Zinn, and other enthusiastic researchers are boldly trying to apply concepts from Buddhist philosophy, particularly Buddhist theories of mind, to scientific topics and have undertaken a deceptively difficult task. In order to apply discipline A's concerns and/or ontology to discipline B's problems, you have to argue for the relevance of discipline B. The usual way of going about this is to note some relevant similarities between A and B such that one speaks to the other. This is particularly a difficult task when the two disciplines don't share many common terms. Such efforts commonly run into problems anticipated by philosophers of science, particularly Kuhnean incommensurability. Interdisciplinary work between Buddhist philosophy of mind and Cognitive Science has an especially acute need for clarity on these matters, since the disciplines in question don't overlap as much as interdisciplinary work between sciences. Jose Ignacio Cabezon, in a footnote to his chapter in Wallace's edited volume Buddhism and Science calls explicitly for more engagement by philosophers of science.

> "One might say that expertise in the various sciences themselves is already well represented in the dialogue but that what has been missing from the side of science is a perspective with a broad overview of science, as exemplified in disciplines like the history, sociology, and philosophy of science. If I have a suggestion to make in regard to science, analogous to the one I put forward with respect to Buddhism, therefore, it is that representatives of these latter disciplines be

brought into the discussion in a more consistent and self-conscious manner." (Wallace, 2007, p. 65)

In this chapter I'll apply the model of analogy I discussed in chapter 3 to the problems of Buddhism and Science. I'll explain how analogical arguments rely on judgments of basic physical similarities between elements within fact-pairs and the relevance of those elements to the question under study.

Artificial Intelligence is a branch of computer science which tries to replicate the functions (if not the substance) of living beings with minds, especially replicating some functions of human thought. This focus sometimes brings together AI researchers with other fields of Cognitive Science, in that new information about how brains work will change the software requirements which explain different processes an AI system may be designed to mimic. I will be relying on a version of the WG-A model which includes space for negative analogies (See figure 9) which explicitly represent relevant ways in which the two scenarios are dissimilar. Representing dissimilarities is essential to being able to judge the plausibility of a resultant analogy.

Dissimilarities always abound in analogical reasoning. After all, the point of an analogical argument is to show some relevant similarities in separate domains which are otherwise dissimilar. In order for analogical reasoning to function, it's necessary to determine the difference between relevant and irrelevant dissimilarities. The difficulty this task causes is highly relevant to these interdisciplinary arguments. Relevant differences ought to make an analogy less plausible, but it can be difficult to make a solid case for the relevance of a difference.

The importance of this element of analogical arguments is highlighted by Walton, who points out that the relevant dissimilarities are crucial in most methods of evaluating the strength of an analogy. Walton (2014) The information about the relevant dissimilarities between the objects compared must not only exist at the point where an analogy is originally proposed, but must also be used in case a rival analogy is later proposed to explain the same

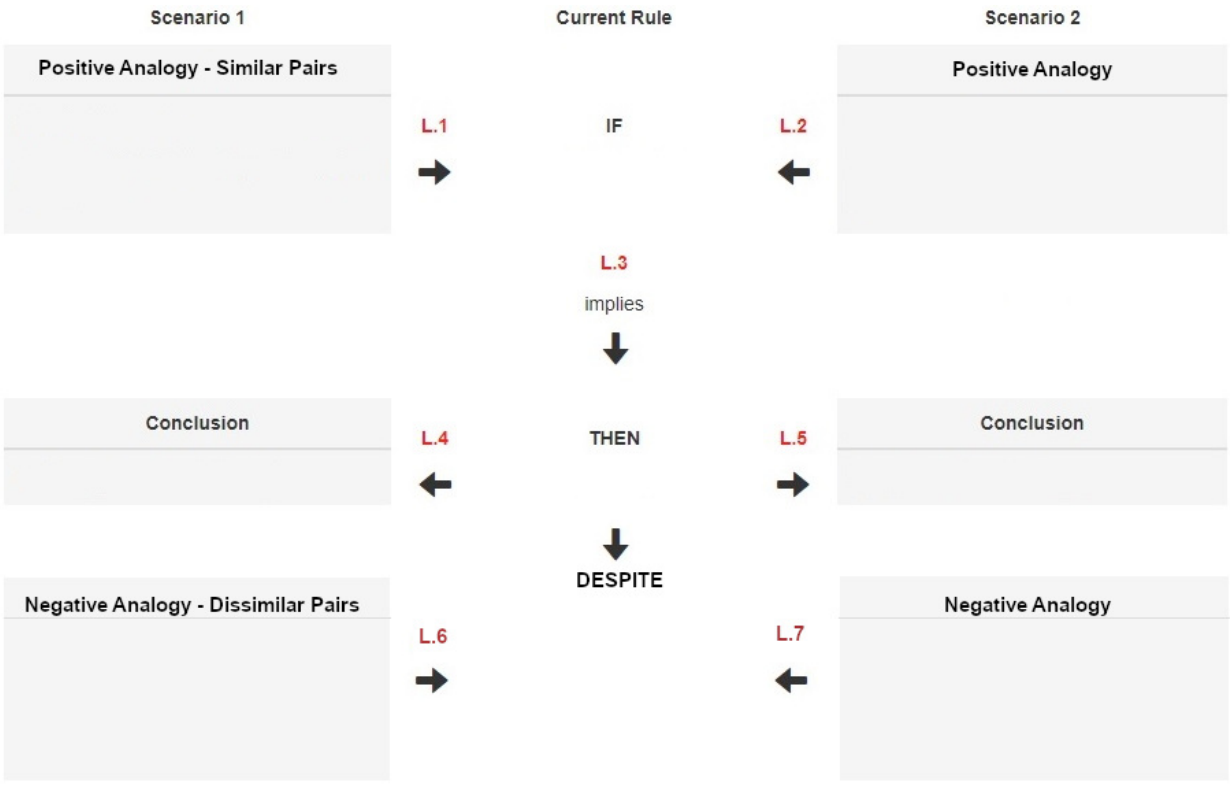

Figure 4.1: WG-A With Negative Analogy

phenomenon. In that case, the original analogy's relevant dissimilarities would be a relevant factor in determining which rival model best fits observed evidence.

The trickiness of this kind of argument is apparent in B. Alan Wallace's Taboo of Subjectivity. Here, he is trying to construct a counterargument to reductionist accounts of mind. He launches a criticism of an argument for reduction based on an analogical argument, and finds that it is inappropriate to compare brain activity and mental activity. He argues that the frame of a "low-level" physical analysis of brain matter cannot be compared to "high-level" complex events. According to him, such a low-level frame can never be relevant to arguments about high-level matters. To believe that, however, is to dramatically underestimate the power of analogical arguments, which are specifically useful because they compare things in different contextual frames. The different contextual frames are not reason enough, therefore, to dismiss an analogy.

> To raise a counterargument in defense of the emergent status of the mind from matter, one could point out that the fluidity of water is indeed a classic example of an emergent property, but it is a primitive one in comparison to the emergence of simple behavior such as an insect's or a robot's ability to walk. Thus, the real dissimilarity between the emergent status of fluidity in water and the emergent status of consciousness from the brain is that the former is a low-level, or primitive, emergence, while the latter is a high-level, or complex, emergence.(Wallace, 2004, p. 137)

In addition to this failed analogical argument, though, he provides a more plausible reason to doubt reductionism, albiet, one that doesn't rely on the workings of analogy. He says that in order to prove the emergence view, we'd need a much better picture of the internal workings of identified brain regions, not just their gross interactions. Absent this, we have no way to rule out a causal role for a non-material mind. Though I have no qualms with this argument as such, I would argue that the proper response to the lack of compelling arguments ought to lead to agnosticism as to the nature of the ultimate material of the mind, a result that Wallace opposes.

4.2 Analogies, Expertise, and Fixing Definitions

On my analogy model, Kuhn's incommensurability looks like a variety of equivocation over a term which appears in paired statements of different domains being compared. In order to fix the meaning of the term with more precision, it's precise meaning in different disciplines requires at least a basic understanding of the associations and relationships that help to determine the term's meaning in each discipline. It is not enough to simply learn a dictionary definition for a term in order to apply it with the precision necessary to avoid equivocation. This reasoning calls for active engagement with Buddhist scholars, as they have the best qualifications to judge their discipline holistically.

As a result, this abstract, philosophical notion of word meaning changes our picture of proper interdisciplinary engagement. When someone like Kabat-Zinn tries to use a Buddhist notion without engaging with the rest of Buddhist doctrine, they run the risk of having a crucial term fall out of step with the term in the opposite half of the fact-pair. Jose Ignacio Cabezon worried over some of the dangers of leaving Buddhist scholars out of the conversation in Wallace's Buddhism and Science thus:

> There is, of course, an inherent danger in the scientific objectification of subjects in an experimental setting. The peril lies in the possibility that those being tested come to be considered mere objects and thus dehumanized. Such a problem becomes especially acute when subjects are separated from researcher not only by professional but also by cultural distance. One way to lessen the negative effects of scientific experimental objectification... is to involve the subjects, to whatever extent possible, in the actual planning and execution of experiments, that is, to acknowledge them as intellectual equals and thus to give them a voice as colleagues. (Wallace, 2007, p. 37)

When the exact meaning of a term on the Buddhist side is in question, scientists should defer to experts in Buddhism. This requires live interdisciplinary interactions between subject matter experts. Failure to properly engage with these scholars not only risks the common problems of dehumanization of your object of study, but also results in studies with diminished *scientific worth* by shutting down analogical arguments which might otherwise yield interesting and useful similarities.

Ronald Purser's criticism of Kabat-Zinn's mindfulness movement is prominent in scholarly circles. In an influential critique of Kabat-Zinn, Ronald Purser argues that the operational definition of mindfulness that Kabat-Zinn uses in his MBSR practice focuses on deliberate attention to the present moment is not sufficient to explain all of the aspects of Buddhist mindfulness. Purser (2015) Purser's central criticism is that Kabat-Zinn has created an excessively shallow focus on physical pain, anxiety, stress, and depression, but fails

to fully appreciate the complexity of the term in Buddhism. In simplifying the definition of the term, Purser argues that Kabat-Zinn leaves out many of the important aspects of mindfulness that are essential for addressing the suffering caused by fear of change, and the suffering caused by the illusion of a permanent and unchanging self. Purser identifies these three senses of the word *dukkha*, which is often translated as "suffering," which are all invoked when the word is used in a Buddhist context. "I propose that MBSR's present-moment focused operational definition limits the depth and potential of secular mindfulness practice to further investigate the temporal structure of suffering, or dukkha, at a fundamental level. In the Buddhist teachings, there are three forms or levels of suffering..." (Purser, 2015, p. 680) He claims that Kabat-Zinn's approach to mindfulness meditation only addresses the first, and most shallow form of suffering. Since Purser's argument is that the Buddhist term translated as suffering is substantially broader than the way it's used in Kabat-Zinn's MBSR.

Purser argues that this definitional slippage between Buddhist *dukkha* and Suffering as defined by MBSR is the result of two concepts which are missing in Kabat-Zinn's treatment of the subject: the doctrine of no-self (*anatta*) and an understanding of process ontology. Understanding both of these doctrines is necessary to the two subtler forms of suffering that Purser identifies. He argues that the lack of these important concepts deprives the term "suffering" as Kabat-Zinn uses it in his MBSR practice of much, though not all, of it's therapeutic usefulness. Since Kabat-Zinn is explicitly concerned with the efficacy of his system at addressing suffering, this criticism indicates that Kabat-Zinn's practice may be falling short of it's full potential as judged by the standards Kabat-Zinn set himself. Since Kabat-Zinn avoids deep discussion of fear of change, and the suffering caused by the illusion of self, meditators trained in the MBSR tradition will have difficulty understanding these forms of suffering. Purser identifies the MBSR version of suffering with the first and shallowest of the three senses of the word. In order to appreciate the fear of change and overcome it, it's necessary to understand Buddhist positions on process ontology and codependent arising. The fear of change is an essential part of understanding what Buddhists mean by

dukkha in the first noble truth. The first noble truth in Buddhism expresses the idea that life is inescapably unsatisfactory, even when you're experiencing great success. This suffering comes, in part, from fearing that the wonderful things you enjoy when you're experiencing success are often temporary, so fearing to lose those wonderful things is a natural response. This fear is an important part of *dukkha*, but not one that Kabat-Zinn can address with his "operational" definition of suffering, which focuses on valuing the present moment. As Purser points out, Kabat-Zinn's focus on the present moment not only misses this point, but in fact could exascerbate this kind of suffering by enhancing our attachment to the present moment.(Purser, 2015, p. 682)

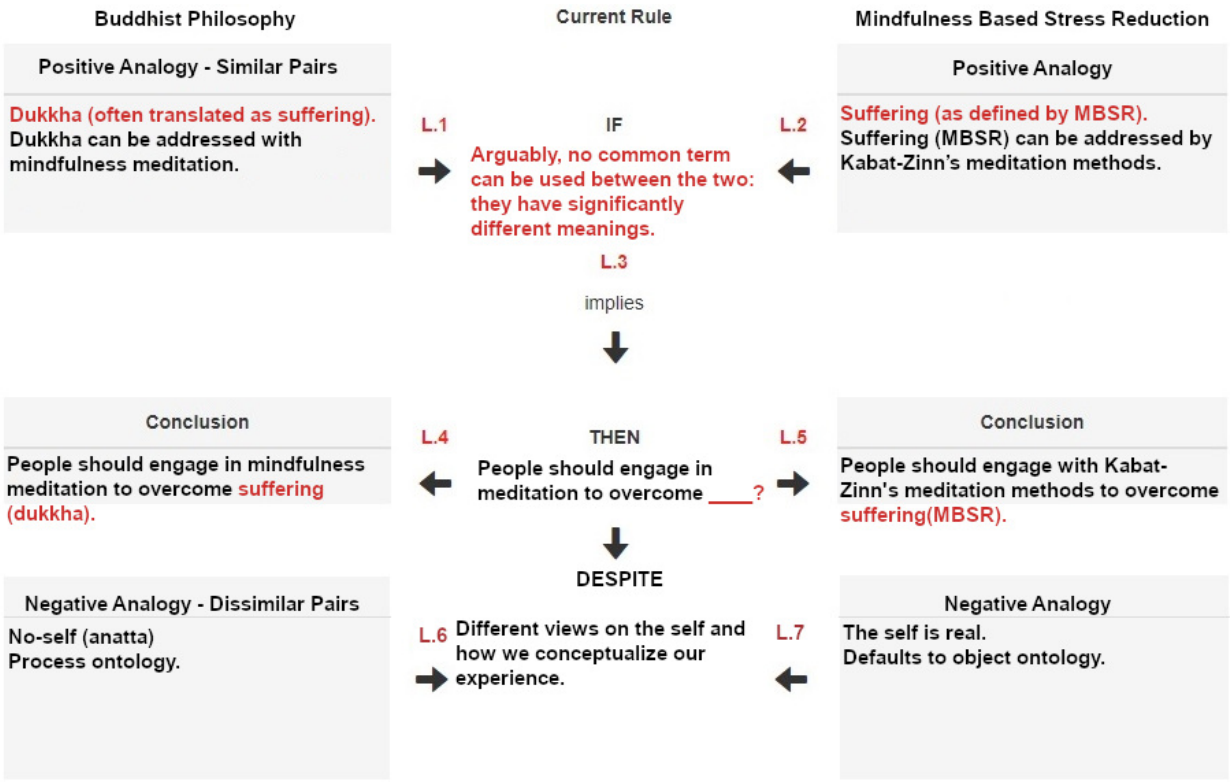

Figure 4.2: Model of Zabat-Zinn along the lines of Purser

Despite leveling this serious criticism, it's remarkable that even Kabat-Zinn's fiercest critics admit that, at least on the level of superficial physical and emotional pain, MBSR techniques have garnered significant scientific support. Even Kabat-Zinn's fiercest critics

must admit, however, that there is good empirical support for the contention that MBSR actually does relieve stress and it's associated suffering. "The practice of cultivating present moment awareness, as much of the research on MBSR suggests, has demonstrated therapeutic value for reducing stress and a variety of other symptoms fueled by excessive rumination." (Purser, 2015, p. 680) So, from the perspective of the relief of suffering, which is the Buddhist soteriological goal, it's difficult to be too critical of Kabat-Zinn. The most pessimistic critics worry that by addressing shallow suffering without giving his students the tools to deal with deeper forms of *dukkha,* Kabat-Zinn might create an addictive (many of the studies show differences in psychoactive chemicals like serotonin in the minds of meditators) form of meditation practice that fails to address more serious suffering and prevents practitioners from seeking help with that deeper suffering.

With a concrete model of analogy, it's easier to explain Purser's response to Kabat-Zinn's work. An analogy depends, fundamentally, on the similarities represented by fact-pairs in the two scenarios. In this case, though, Purser points out that Kabat-Zinn uses a key term (suffering) in a way that is inconsistent with the traditional Buddhist term *dukkha.* The effect of these two different definitions is to relevantly limit the similarities between the fact pairs given in the positive analogy. That being the case, any generalizations made over these relevantly different terms is highly suspect. This has the double-effect of making the analogy less plausible by adding to the negative analogy section, but also limits the ability of the model to generate a Toulmin-style warrant (the center column). I have highlighted worrisome aspects of this potential analogy in figure 10 in red. As I've argued, analogical reasoning provides a fundamental structure that allows knowledge to cross disciplinary boundaries, but the process might be halted by subtle differences in word meaning across the domains under comparison. Failing to use the word in the same way in each instance casts serious doubt on any generalities we might draw from the analogy (represented by the central column in the figure.) I argue that this kind of shifting definition demands that people engaged in interdisciplinary scholarship with Buddhist philosophy maintain strong interdisciplinary ties

to ensure that key technical terms are being used in a way that is consistent across domains. This is not just a demand of polite professional association, but a logical necessity for an analogical argument to proceed.

4.3 Problems that NLP is Struggling to Solve

One of the research lines in AI, which shows promise in representing the (sometimes irrational) associations which help determine word meanings, came in the form of vector-based word representation. This research program had a famous early success in the program Word2Vec, first published in 2013. Word2Vec was designed to fulfill the requirements of John Rupert Firth's distributional semantics, which was specifically designed with Wittgenstein representations of context and holism in mind. Skelac and Jandric (2020) As a result, creating word representations robust enough to support analogy is a common problem in models of language in this research line. A famous example from Word2Vec, one of its early successes, was in building a system that could answer the question "Man is to woman as king is to ?." However, this system failed to provide correct answers to other analogy-based questions and failed to offer any comprehensible explanations for it's failures and successes.

Systems of vector-based word representation have continued to develop in the years since, and one of its most lauded successors is named GPT-3. "GPT-3 ("Generative Pre-Trained Transformer 3) is a language model — a program that is, given an input text, trained to predict the next word or words. GPT-3 is one of the largest such models, having been trained on about 45 terabytes of text data, taken from thousands of web sites such as Wikipedia, plus online books and many other sources." Mitchell (2020) So, in the fast-moving realm of AI research, GPT-3 is one of the best candidates for representing words well enough to adequately address family-resemblance problems. It's network of language representation is based on the associations between different words that is formed by observing millions of human uses of the word. It is well-positioned to address the problems fixing natural-kind

terms that Kuhn worried about becoming incommensurable. As a result, it's also the most promising research track to address the ongoing problem of fragile systems.

Fragile AI Systems Analogy is seen as one of the major hurdles in developing more general, flexible forms of AI. For example, one memorable presentation the Association for the Advancement of Artificial Intelligence conference told a story about a robot which was programmed to use a spatula, successfully. When the researchers tried to teach it to use a different model of spatula, however, the robot was completely incapable. The problem was teaching the AI program operating the robot to recognize that one tool was similar to another. Problems like this in AI systems are common, and the term that researchers use for that is to say that the resultant algorithms are fragile. Lee (2020)

The problem of fragile AI systems is not just an abstraction for researchers to be concerned about, but in fact have life or death consequences for devices that are currently in operation. A recent interview of Raj Rajkumar, an electrical and computer engineering professor at Carnegie Mellon University in Slate magazine, discussed this problem of fragile AI systems in relationship to a series of car accidents in which a Tesla car operating under it's autopilot system struck a stationary emergency vehicle with it's lights flashing. The National Highway Traffic Safety Administration has claimed that such accidents were responsible for 17 injuries and one fatality between 2018 and August 2021, when they opened an investigation. Raj traces this problem to the autopilot system not being able to determine that emergency vehicles, seen from the side and with flashing lights, were not explicitly represented in the "hundreds of thousands, if not millions" of images of vehicles that were used to train the AI system. Like the spaula-using robot, this AI system was unable to determine that the new images of emergency vehicles were similar to the numerous examples it was trained upon. Mak (2021)

One reason this problem receives a lot of scrutiny within Cognitive Science is that these fragile systems need with far more instances than a human needs in order to respond rea-

sonably to stimuli. They are easily thrown off by novel inputs because they cannot judge similarity to those inputs they have already received. Such systems don't understand concepts like "car" and "spatula" the way we do, since they are not able to add novel inputs to the natural-kind groups based on similarity to what they have seen. This is precisely the problem that Kuhn was concerned with when he examined Wittgenstein's family resemblance problem. (Kuhn and Wilson, 2001, p. 200-2) The root of the issue may be that these words are not being represented with the kind of richness of data which would allow similarity judgments to go forward. The problem of Kuhnean incommensurability and the problems of spatula or car-operating robots revolve around the same fundamental issue: how can we represent the meanings of conventionally-useful objects in such a way that will allow robust judgments of similarity to proceed?

Robust similarity judgments call for an organized model of analogies which will accommodate simple, physical similarities and have the possibility to scale up to extremely complicated judgments of similarity which might accommodate judgments of similarity between scientific theories. Despite notable early successes, modern efforts in NLP have fallen short of this goal. "All in all, GPT-3's performance is often impressive and surprising, but it is also similar to a lot of what we see in today's state-of-the-art AI systems: impressive, intelligent-seeming performance interspersed with unhumanlike errors, plus no transparency as to why it performs well or makes certain errors. And it is often hard to tell if the system has actually learned the concept we are trying to teach it. I don't have a definitive answer for the question in the title — can GPT-3 make analogies?" Mitchell (2020)

Researchers in AI fields which find themselves struggling to represent natural kinds in the way that humans do would benefit from additional philosophical discourses that explore problems related to these conventional categories of language, how object-based conventions differ from a process-based representations of reality, and what different mental abilities are invoked or used in the formation of these ideas.

4.4 Buddhism Speaks to these Problems

Buddhist philosophy of mind is ideally suited to address many of the problems that researchers in NLP are struggling with. In their devotion to uncovering cognitive illusions, scholars of this period took pains to describe the necessary connections between concepts, and to provide a framework for establishing the minimum requirements for a cognition to be conventionally useful.

Buddhism's discussion of conventionally useful objects in language argues that our minds are primed by their nature and circumstances to build a useful object-based ontology to help us understand the world and note where this object-based ontology fails to represent the world well. When our object-based understanding of the world falls short, they suggest alternative concepts which make use of a process-based understanding of the world. In artificial intelligence, there are several kinds of systems being developed which bear striking similarities to these ontological systems. AI researchers working towards general (flexible) intelligence are working on physics engines which help a robot understand the physical world and models of affordances, which catalogue and try to recognize opportunities to use objects in relevant ways. Both of these systems are more heavily process-oriented than models in NLP which try to represent word meaning. Debates in Buddhist Philosophy that concern whether something is a conventionally useful label in an object-based ontology or should be understood through process provide interesting insights into this issue.

One problem that AI researchers have struggled with is creating "resource rational" systems to solve intellectual problems. In the case of Tesla cars hitting emergency vehicles, I noted that their AI system needed to be trained on hundreds of thousands of pictures of vehicles, and might need additional pictures to resolve their tendency to crash in specific situations. This is a particular source of frustration for AI researchers because human (and animal) children do not seem to need so many paradigm cases to be presented as examples for them to recognize emergency vehicles. The architecture that the program uses to make it's decisions is not as efficient in its resource needs as minds seem to be. Problems

in the world often become computationally inefficient when trying to compute affordances, because, among other reasons, there are too many potential abstractions to grind through them all. This problem, observed by AI researchers, goes a long way towards explaining why we create the conventions we do, despite the fact that they frequently lead to mistakes and illusions. Buddhist insistence on the mere conventionality of language and simultaneous acknowledgement that we cannot do without our conventional categories is entirely reasonable if our minds, like AI algorithms, are not capable (at least without training) of producing the intellectual power (compute cycles) to conceptualize everything as a process.

The network that represents these meanings must therefore be able to represent and differentiate causal forces which relate to the dharmas. This points to a weakness of GPT-3: in it, the associations between words are all represented by the same kind of vector, a kind of association that sometimes stands in for similarity. One way that Buddhist philosophy might contribute to the building of such models is to indicate which kinds of cognitions rely on judgments of logical or causal entailment, as opposed to mere association. The needs uncovered by that search reveals additional needs in word representation which might not be adequately met by current efforts.

The problem of explaining the decisions of AI programs is a significant one, and it's one that might be aided by a careful study of Buddhist models of mind. One of the techniques in AI development is called machine learning (ML). ML systems are programs that are given examples of ideal performance on a task and use repeated experience to learn how best to approximate ideal performance on it's own. ML can be further subdivided into supervised, unsupervised, and reinforcement learning systems based on the amount and type of human input required to guide programs towards useful procedures. In all of its varieties, ML aims to have a program teach itself how to perform the task, which is only described in terms of it's end goals. Bringsjord and Govindarajulu (2020) Since the writers of ML algorithms are not working out the exact procedure to accomplish the program's goals, one troublesome feature of ML algorithms is that researchers usually have a hard time knowing why these

programs either work or fail to work. I argue that this explanatory gap can be addressed, to some extent, by different theories of how our minds conceptualize our experiences in pursuit of our own, human, goals.

The idea to look to theories of human cognition to explain possible or ideal AI procedures has a long history. One of the methods that ML researchers use to help their algorithms "learn" is through the use of data structures modeled on networks of neurons in a human brain. They call these structures "artificial neural networks," and frequently go so far as to drop the word artificial in discussions of the term. Bringsjord and Govindarajulu (2020) Researchers using this technique specify inputs and desired outputs, but use a blind, algorithmic mechanism to generate the layers of analysis that helps you get from input to output. The neurons of these models are individual processes which take inputs from either experience or a neuron on a different level of the data structure. This lack of explanation creates interesting opportunities for interdisciplinary theories to fill the gaps.

Process and Dharmas in AI Much of western metaphysics is predicated on the assumption of an ontology of objects. Through an explanation of all the types of objects there are, some thought to develop a method of categorizing all of the things in the world. This foundation set the conditions that caused Heraclitus to worry that a person cannot step into the same river twice. By analyzing a river as an object, we may overlook the aspects of a river that result from it's motion, that is to say, it's changes over time. As Heraclitus' example makes clear, the ontology of objects doesn't work very well when applied to something that changes swiftly, like a river. A later figure in the Ancient Greek world, Aristotle, firmly set western ontology on object-based grounds, arguing that the essence of an object must be understood through characterizing it's substance, and not by characterizing the way it changes. The paradox of Heraclitus' river cannot be resolved within a purely object-oriented framework, however, so things that change quickly over human-relevant timescales, like mental processes and rivers are not well explained within this framework.

A related doctrine from Buddhism which is very relevant to a live problem in AI is how to generate representations of linguistic entities with duration in time. In an intriguing presentation at CogSci 2021, Zhou and Yurovsky (2021) the Carnegie Mellon Psyhologists Zhou and Yurovsky noted that children learn nouns more easily than they do verbs, and this was true regardless of the language that they spoke. So, they set out to study how unsupervised learning models in AI approach the problem. One question they had while working with these models is how to teach the models the meaning of a verb. The unsupervised learning models they were working with used pictures as inputs,Kuznetsova et al. (2020) but noted that this static picture doesn't necessarily depict verbs well, which include changes over time. In exploring alternatives to this issue, they have begun to follow a protocol from MIT which analyzes video into multiple snapshots over time (16x second) to form ordered datasets of photos to use in ML program training. They argue that by taking these snapshots, an unsupervised AI learning model may be able to discover patterns in the changes observed from snapshot to snapshot. The patterns that would provide hints as to the way physical objects change over time, though the information gathered is based on data on a series of static moments. Though I don't claim that Buddhism offers a univocal picture of representing time, there are many interesting observations in Buddhism relating to the issue. As Ronkin points out, Buddhist philosophies offer treatments of time which bear some striking similarities to the efforts at MIT. It stands to reason that the Buddhist arguments which have so many relevant similarities might be a fruitful place to find apt metaphors which would help to guide future research.

> Dhammas as the cognitive objects of manovinnana qua mental cognitive awareness may now be better rendered as apperceptions in the sense of rapid mental events by means of which the mind unites and assimilates a particular perception, especially one newly presented, to a larger set or mass of ideas already possessed, thus comprehending and conceptualizing it... capacities or capabilities of psycho-physical events: short-lived minds or consciousness-types (citta)

> that interact with material phenomena, each of which arising and ceasing in sequential series while having its own function and capability. (Ronkin, 2005, p. 40)

This approach to time analyzes events or processes in time as a series of moments, each of which must contain its own atomic dharmas, representing irreducible bits of our mental experience. Each of these dharmas persists for only a moment, but they may cause the existence of similar dharmas to occur in the next moment, thus often giving the illusion of persistence. Things we do which seem to have effects later in time are, under this analysis, metaphorical seeds which are planted to sprout later. The researchers working on MIT's video analysis would have to follow similar mental arguments to get their algorithm to represent what changes occur between the 16 frames per second, and which elements persist. These similarities might lead to interesting new observations, when used in a scientifically acceptable analogy.

Anatta (No self) and Neural Models The rejection of object-based ontologies was, in part, about rejecting the idea that a thing can be individuated based on the essential features of the substance from which it is made. This idea is also behind the Buddhist rejection of a view of the self which relies on essential, unchanging properties to individuate a person through time. This rejection of an object-metaphysics of selfhood has interesting applications to AI in it's own right. Another interesting area where Buddhism speaks to live research problems in Cognitive Science is in assessing high-level structural models of cognition. In order to model the mind well, we need to be able to account for meta-cognition. The ability to know ones-self is, after all, one of the key elements that makes human thought special. In high-level structural models of cognition, this creates a challenge, to adequately describe the interactions of different mental modules. In an interesting study concerning the mind at rest, three psychologists from the University of Washington, led by Catherine Sibert discusses prominent high-level models of cognition and tries to associate them with specific

brain regions for study.Sibert et al. (2021) In running their study, Sibert et al found that the brain regions identified with each of the mental functions under study were broad and overlapping. For the purposes of their study judging the interactions between these brain regions, they focused on the largest, non-overlapping clusters of brain activity associated with those regions. They were trying to find out if any of a number of common high-level models of cognition accurately predicted the brain regions that would activate (under fMRI study) in goal-directed and mental rest state trials.

The Hub and Spoke structure that Sibert's study examined includes Common Model of Cognition, the model which performed best in the study. It's clear, in her description of these models, that the mental module in the hub position is too central to be consistent with the *anatta* doctrine. "In the 'Hub-and-Spoke' family (Fig. 1C), a single ROI is designated as the central 'Hub', and is bidirectionally connected to all other ROIs. However, none of the 'Spoke' ROIs are connected to any other - all activity must travel through the "Hub"." (Sibert et al., 2021, p. 2) On the other hand, hierarchical models don't fare any better on this score. "The 'Hierarchical' family of models proposes an alternate structure, wherein brain connectivity implements hierarchical levels of processing that initiate with Perception and culminate with Action. Networks in this family conceptualize the brain as a feedforward neural network model in which different regions perform progressively greater levels of representational abstraction." (Sibert et al., 2021, p. 2) In arranging mental modules into a hierarchy, it's tempting to label the pinnacle of the hierarchy a kind of higher self. That would be opposed by the *anatta* doctrine.

The Buddhist doctrine of *anatta* has some interesting things to say about modeling the mind at this level. In the thread and cloth example, proposed by Vasubandhu and discussed in chapter 2 of this work, it was argued that there is no part of a piece of cloth, no special thread whose nature determines the essence of the cloth. In doing so he was making a point about how our minds create conventionally useful aggregations of objects, like numerous threads woven together. It's not a particular thread that determines the nature of the cloth,

but a process of judging in our minds which determines the overall nature of the cloth. He was making this argument, in part, to argue something similar about the conventional notion of the self. He was arguing that there is no core of our minds which is special, central, and determines the nature of the whole. In doing so, it created interesting arguments which could apply to choosing a high-level model of cognition. While Sibert's study discussed the analogy of a wheel (for it's hub-and-spoke model) and the model of human organizational structures (hierarchical model), Buddhism offers it's own, alternative, analogy to offer it's own proposed mental structure.

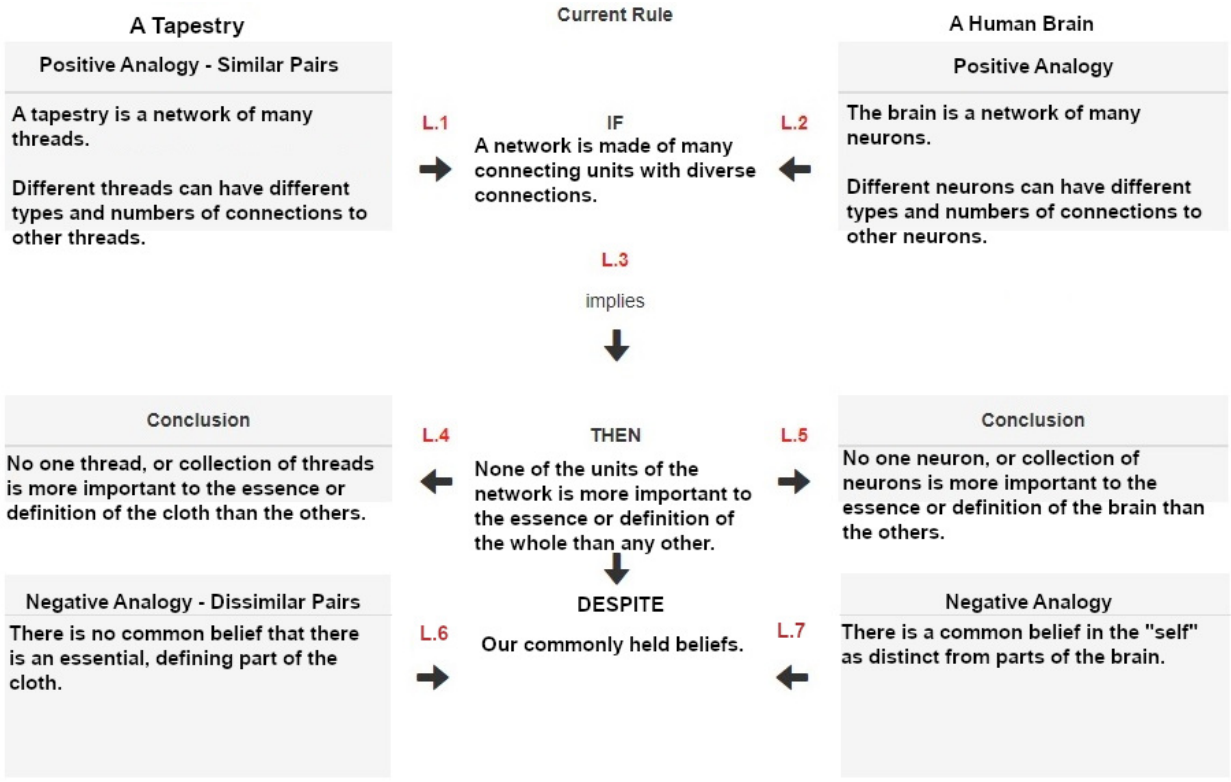

Figure 4.3: Analogy Modeled as a Tapestry Compared with a Mind

While Sibert et al treat the different parts of the CMC model as if they are separate modules, they had difficulty, from the start, in differentiating those modules. One of the first obstacles their study design had to overcome was the overlapping nature of the regions they identified in the brain which activated when different mental functions were used. It

is as if these regions are objects which pass information from one to the other. On the other hand, Buddhism, in it's focus on process, identifies different mental capacities, and tries to show how dharmas are heavily dependent on their connections for their meanings. This would plausibly result in complex feedback loops between mechanisms which would require complex connections. If those dharmas have any correlated physical state in the brain, we should expect them to connect with these different brain regions. As a result, we should expect fine and complex connections between the physical corollaries of these mental functions, and that is exactly what Sibert et al. report. The interdependence of dharmas, therefore, might tell us some interesting things about the way that nerve clusters interact with one another, by offering commentary on these models.

The *anatta* doctrine suggests that neither a Hub-and-Spoke nor a Hierarchical model will be sufficient. Either option seems to grant one brain region special status, either by putting it at the top of the hierarchy or by placing it in the "hub" region of the Hub-and-Spoke. The Buddhist doctrine, in it's exhaustive discussion of self-hood, argued that there was no part of the mind which is central in these ways. They argue that the illusion of self is partly based in (often inappropriately) thinking of the self in object-terms, rather than grappling with the additional complexity that process considerations add.

4.5 The Robot's Challenge

A professor's helper robot, tasked with retrieving a red pen from a closet, would not be terribly helpful, given the current state of artificial intelligence. The command, which is perfectly comprehensible to a human helper, "Bring me the red thing" is vague and implies that there is only one red thing. A human given these options (see figure 10) would be able to determine which of them seems more red in the context in which the command was given. A partially ordered set, transformed by the context of the phrase, might produce a single dominant case of redness even if multiple things fall into the red category. An unordered set would give no indication which member of the set should be chosen first, so that model

seems inadequate. An ordered set would return the same result regardless of the context. It's only by using a partially-ordered set whose order is sensitive to context that we can solve this problem.

In order for a robot helper to behave the way that a human helper might, it's necessary to have a representation of what the various options' associations. This result is supported by Buddhist arguments about codependent arising. This doctrine points towards the crucial role of context in discerning conventionally useful objects. In the robot's case, the context would rule out the historical examples given, such as a picture of Chairman Mao. The red-communist association is not relevant here, and that would be clear from a lack of dependency on words associated with communism.

One way that a robot programmed with Buddhist philosophy in mind might help is in its analysis of creating conventionally useful objects out of an aggregate. Seeing the pen as an aggregate with meaningful parts could prevent the robot's error of only bringing the pen cap instead of the whole pen. Another way the robot's performance may be improved is to arm the robot with the associations that people draw upon in recognizing different objects. For example, the robot may fail if the desired object is concealed within a box. However, a connectionist model may highlight this possibility by recognizing associated terms on the box (i.e. Bold, fine, rolling-ball, ink, precise, Bic, Pilot, etc. indicate the likelihood of pens within.) These words or phrases don't need to be a part of the necessary and sufficient conditions which are traditionally used to define sets in symbolic representations of word meaning. Resolving these practical problems would require a language of word representation that is capable of extreme flexibility, relative to context.

Artificial intelligence researchers working on natural language processing are beset by serious philosophical problems. These problems are best understood through the philosophy of Thomas Kuhn, who noticed that it leads serious difficulties when two domains are using terms in subtly different ways. Despite the difficulty, fruitful interdisciplinary comparisons are still made. Through the use of carefully managed comparisons, especially as they relate to

overarching models of the self or conceptualizations of time, it's possible to make comparisons that bear on contemporary problems in AI research. Researchers concerned with improving their solutions to these problems would be wise to consult these ancients, and the work of scholars who have analyzed and criticized this work over the years.

Chapter 5: Limitations of the Analogical Approach

I have argued for a model of analogy that depends on plausibility, but this is not an uncontroversial choice. Plausibility is difficult to apply, relative to the alternatives that deductive and inductive models offer. A successful deductive model would likely offer concrete answers to whether or not an analogical argument is a good one, but it's difficult to make such a model contextually sensitive. A successful inductive model of analogy would likely offer a concrete, numerical probability that a given analogy is a good one. This numerical analogy score could then be used to weigh different analogical arguments against one another to see which one is preferable. Unfortunately, that numerical score would not be sufficient to compare two rival analogies, given the complex trade-offs involved in judging desiderata which would need to be considered when comparing analogies. Though it would be convenient to have a successful model on either strategy, as they would yield simple results that are easy to work with, neither seems to plausibly model the way that we use analogies.

I have endorsed a model of analogy based on defeasible-reasoning, which argues that analogical arguments, including ones in the sciences, yield only a judgment of plausibility, to be judged on the basis of multiple desiderata. Plausibility is controversial because it can be decided many different ways, even if we agree on a single set of desiderata. If we are to use plausibility long-term, the rationale behind the desiderata's evaluations must be preserved, so that it can be compared to other analogies.

This analogical method for bringing different disciplines into communication with one another is contrary to the assertions of Gould and Wallace, who argue for a strong separation between different disciplines. Gould argues for this point on the basis of apparently separate goals pursued by those who study the introspective methods of Buddhism and those who

study the mind (in their various specialities) in science. That argument seems weak when we're discussing the study of the mind, since the Buddhist goals of reducing mental suffering is entirely consistent with western psychological goals.

Wallace, on the other hand, argued that these areas couldn't come into fruitful communication unless they are both built on a common beliefs about the immaterial makeup of important mental causes. On the contrary, working out an explicit model of analogy only emphasizes the fact that the ultimate microscopic material causes for things are rarely relevant to an analogical argument, which often focuses on higher level isomorphisms without any assumption that the ultimate physical components on both sides of the analogy will be the same. AI researchers are often explicitly looking for functions which can be replicated by silicone chips, rather than meat. There is no reason to assume that AI researchers and Buddhists need to agree about the nature of microscopic causal components of the mind in order to speak fruitfully.

5.1 Difficult Foundational Needs Required

Plausibility requires high demands on us. It's judgments only makes sense relative to a holistically understood foundation. Bartha draws particular attention to predictiveness, applicability, scope, and simplicity in their role in evaluating analogical arguments. Scope, in particular, suggests that the best theory has wide applicability, but simplicity suggests that the best theory is the one with the fewest components. Not only can these desiderata conflict, but they both must be judged against the basis of established theory. In order to use Bartha's model then, we must also rely on a representation of other ideas which lay beyond the model. The need to articulate a model of these other ideas would be necessary for any implementation of analogical reasoning in artificial intelligence.

It's difficult to conduct research into interdisciplinary areas of scholarship. Among the difficulties is a prominent worry that different fields will be using the same terms in incommensurable ways. In my analysis of Kuhn's arguments about incommensurability I found

that the difficulty he anticipates relies heavily on a view of word representation that uses a word's position in a network of associations to determine it's meaning. Without this fine-grained representation of meaning, it's hard to make effective analogical arguments for fear of equivocation over the meaning of a key term. This is the root of the difficult issue of holism in Kuhnean judgments, which is a particular impediment to interdisciplinary work which relies heavily on analogical arguments.

5.2 Analogical Methods Compared

The path to a more effective interdisciplinary crossover between Buddhist scholars and cognitive scientists runs through analogical reasoning, therefore. It's in the comparisons between starkly different frames of reference that worries about Kuhnean incommensurability arise, and analogical arguments are the ones that specialize in such comparisons. There is, however, significant disagreement about how these arguments should be analyzed and interpreted. In *By Parallel Reasoning* Bartha (2010), Paul Bartha laid out a convincing case for his model of analogy on the basis of analogical arguments which have served to advance scientific and mathematical judgments. By selecting these cases to model his theory of analogy on, he draws on material that is already widely accepted as successful scientific practice. These serve as paradigm cases of analogical reasoning, and Bartha is credible in his arguments about the general form analogical arguments should take in science.

From this work, it's clear that analogies don't give clear answers from a simple algorithmic procedure. Instead, analogies have to be judged by a set of desiderata, and the considerations that improve things relative to one desirable trait may make things worse in another. Bartha argued that in order to best assess analogies, we need to balance contrary impulses towards explicitness and economy. He describes this as if there were an advocate and a critic in polite conflict over the structure of the analogy. The advocate would try to complete the analogy as succinctly as possible. In doing so, they are still trying to generalize over the two domains to create a successful analogy. The advocate should be wary of inclusions in the argument

which have questionable relevancy. By limiting the available facts in evidence they cut down on potential problems that might arise from their interactions. The role of critic is to focus on the explicitness of the argument. By doing so, they draw out relevant details of the argument looking for problems that might make the overall argument less convincing. The analogical argument can only become plausible by pursuing both of these goals, which are often in conflict. This conflict is ably modeled by assigning adversarial roles to participants.

The backbone of a plausible analogical argument is the similarities shown between facts in each of the separate domains. In the process of constructing an analogical argument from scratch, the first task is to note relevant similarities between the two domains under comparison. This can be handled with the addition of fact-pairs to the analogy model. The similarity between the fact pairs under analysis is often expressed by using the same syntactic phrase to characterize the subject of a sentence, as in the example "Earth has water. Mars has water." in the historic argument about life on Mars. Mars does have water, albeit not in liquid form at the surface. Water is relevant to the existence of life, so this similarity is relevant to the discussion and makes the overall argument more plausible. The similarity between the two facts is established by describing both planets with the phrase "has water" and the words "has" and "water" seem to be used in the same way in both sentences.

Fact pairs which list relevant similarities make up the part of the argument called "positive analogy." It's also important to note any relevant dissimilarities, which make up the "negative analogy". In the case of life on Mars, a relevant dissimilarity would be "Earth has an atmosphere which is rich in oxygen, Mars has an atmosphere which is poor in oxygen." Since oxygen is a factor that is relevant to the existence of life, this is a relevant dissimilarity which makes the analogy less plausible. When a relevant fact is known about one domain but not the other, that fact is fitted into the "neutral analogy" section.

Generating plausibility is the goal of analogical arguments in science, as opposed to deductive reasoning yielding concrete answers as to whether the argument is good or bad or inductive reasoning which would generate a probability of such. The goal of plausible

analogies in science is the result of balancing different desiderata, in an analogical argument. Bartha follows Kuhn in identifying a limited number of criteria on which a theory should be evaluated: it ought to maximize it's predictiveness, applicability, scope, and simplicity. This competition between competing principles explains why the result shouldn't be understood as a definitive proof or a probability. A theory might have complicated trade-offs between these different desiderata, and that information would be lost if reduced to a numerical probability. Further, these plausible judgements are defeasible. One way in which a plausible analogy might be defeated is by showing that a different, stronger analogy holds.

With this model of analogy in place, the danger of incommensurability becomes more concrete. One possible threat to the effectiveness of analogical reasoning is the worry that two different domains might use the same word in ways that are incommensurable with one another. To say that the words are incommensurable is to say that they have different definitions between which makes comparison difficult. As I argued, fact pairs in our WG-A model often show that they are similar through the use of the same syntactic phrases applied to subjects from the different domains. In doing so, it assumes that the use of the same words is a strong indicator of similarity, but what if the two domains have different definitions for the same (syntactic) word or phrase? In that case, the similarity between the two facts may be the result of a confusion between words with meanings whose names just happen to be identical.

However, this worry about incommensurability evidently does not prevent researchers from using interdisciplinary insights to create large, even revolutionary, changes in scientific theory. Even the most strident critics of Kuhnean holism, such as Davidson, insist on the possibility of interdisciplinary arguments as a first premise, then attack Kuhn for allegedly failing to explain this process. In this dissertation, I've tried to give as concrete an answer to this problem as possible, given the evidences provided philosophical models of argument types and by the current state of natural language processing. I argue that the storied revolutionary insights of interdisciplinary work are the product of a defeasible form of analogical reasoning.

5.3 What is Given Up by Using Analogy

The result of holism is to make it difficult or impossible to achieve true synonymy, unless the two words compared share an identical holistic network. It can be difficult to express similarity in such a network. The holistic evaluations generated by such methods would be difficult to summarize without significant loss. Without this concept of synonymy, it may be more difficult to use formal logic to analyze statements made in natural languages. Under this model, the definitions of terms are in constant flux, since terms are defined by their position in the network, small changes to the network will change their definitions. Adding a new member to be covered by the term, for example will fundamentally change the definition of the term defined by the set. This makes Kuhnean holism difficult to deal with using philosophical methods alone.

One objection that must be dispensed with insists that scientists and Buddhists need to agree on a set of basic metaphysically-defined substances in order to make progress. On the contrary, I argue that in order to conduct interdisciplinary research, it's important to begin with a humility about our previously-established ontologies. A regrettable fraction of the interdisciplinary arguments used to date, particularly those of Wallace, focus on the ultimate entities of existence. They insist that in order for a interdiciplinary science which tries to incorporate Buddhist theories of mind in cognitive science research, we must start with agreement about ultimate metaphysical principles. We must confront, they argue, the difference between a physically reductionist universe and a dualist understanding which includes immaterial minds. Wallace argues that being agnostic about these metaphysical principles is not an acceptable answer, so he set himself the task of converting scientists to his view of metaphysics.

Constructing arguments to convince your interlocutors to accept your metaphysical premises on the strength of a deductive argument is a common and accepted method in religious circles, so the temptation to follow this path is reasonable. However, this strategy is less acceptable in the scientific realm, which insists that large changes in theory be the result of

pressure put on the theory by empirical testing. Scientists settle large controversies over theory through carefully defining experiments so that they can gathering data which is relevant to the problem. The collected body of such experiments is thought to be the determining factor which causes disciplines to shift from one theory to another, hopefully becoming more able to predict novel experimental results with increasing accuracy. It is at the heart of the scientific method to insist on this kind of evidence, and Wallace makes a significant mistake in failing to accommodate these methods.

5.4 What Might be Gained by Using Analogy

This mistake is especially apparent when you consider the area of artificial intelligence, in which mental functions are modeled in silicon and metal wires, rather than biological neurons or supernatural matter. In such analogies, it is clearly understood that the mental functions being compared with similar functions carried out by a computer are constructed of different matter. The material used is immaterial to the argument, so Wallace's concern is simply not relevant in many of the situations under evaluation. I argue that it would be better, and more scientific, to begin with assumptions that are agnostic about such large metaphysical issues, to focus on solvable puzzles. It's difficult to say where the inspiration for new solutions and new ways of addressing puzzles in science comes from. I argue that by focusing on only those elements which are deemed relevant to be included in an analogical argument, we can sidestep the thorny issues of metaphysics and make these revolutionary discoveries.

It's common for funding agencies to emphasize interdisciplinary research in their quest to discover revolutionary ideas. Though they don't often explain the mechanism by which interdisciplinary work is to help, they put a good deal of faith in this mechanism. I contend that one of the primary mechanisms by which these interdisciplinary comparisons are made is analogical reasoning. If it is true, as numerous funding agencies imply, that interdisciplinary work is more likely to result in revolutionary advances in the sciences, it behooves us to

discover how this process works, which includes both how the reasoning method is sometimes effective and also how it sometimes fails. A solid theory of interdisciplinary work may guide policymakers in how best to divide limited resources for research. In this dissertation, I have explained a model of interdisciplinary research that can accommodate serious challenges, such as the worry of incommensurable theories. I hope that, despite the limited scope of this work, this work provides some clarity on how this process succeeds or fails.

Bibliography

Agassi, J. (1964). Analogies as generalizations. *Philosophy of Science*, 31(4):351–356.

Agassi, J. (1988). *Analogies Hard and Soft*, pages 401–419. Springer Netherlands, Dordrecht.

Agency, D. A. R. P. (2020). Broad agency announcement hr001121s0010. *DARPA BAA*.

Ainsworth, C. (2003). The stranger within. *The New Scientist*, 180:34–37.

Albahari, M. (2006). *Analytical Buddhism: The Two-Tiered Illusion of Self*. Palgrave-Macmillan, New York, NY.

Andersen, H. (2000). Kuhn's account of family resemblance: A solution to the problem of wide-open texture. *Erkenntnis (1975-)*, 52(3):313–337.

Anderson, J. (2014). *Cognitive Psychology and Its Implications*. Worth Publishers.

Barrett, L. (2018). *The Evolution of Cognition: a 4E Perspective*, pages 719–734. The Oxford Handbook of 4E Cognition. Oxford University Press, New York, NY.

Bartha, P. F. A. (2010). *By Parallel Reasoning*. Oxford University Press, New York.

Beardsmore, R. (1992). The theory of family resemblances. *Philosophical Investigations*, 15(2):131–146.

Bodhi, B. (2011). What does mindfulness really mean? a canonical perspective. *Contemporary Buddhism*, 12:19–39.

Bogen, J. and Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, 97(3):303–352.

Böttcher, Florian; Meisert, A. (2011). Argumentation in science education: A model-based framework. *Science & Education*, 20(2):103–140.

Bressan, D. (2012). Darwin the geologist. *Scientific American*.

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y.-Y., Weber, J., and Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proceedings of the National Academy of Sciences*, 108(50):20254–20259.

Bringsjord, S. and Govindarajulu, N. S. (2020). Artificial Intelligence. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition.

Buddhaghosa, B. and Nanamoli, B. (2020). *The Path of Purification: Visuddhimagga*. Pariyatti Publishing.

Cartwright, N., Cat, J., Fleck, L., and Uebel, T. E. (1996). *Otto Neurath: Philosophy Between Science and Politics*. Cambridge University Press, New York.

Clark, A. and Chalmers, D. J. (1998). The extended mind. *Analysis*, 58(1):7–19.

Committee on Science and Technology (2009). Onr instruction 3966.1a. *House Subcommittee on Space, Science, and Technology*.

Conant, J. B. (1947). *On understanding science; an historical approach / by James B. Conant ..* Terry lectures. Yale university press, New Haven.

Cooper, M., Fields, L., Badilla, M. B., and Licato, J. (2020). Wg-a: A framework for exploring analogical generalization and argumentation. In *CogSci 2020*, pages 1–7.

Coseru, C. (2017). Mind in Indian Buddhist Philosophy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2017 edition.

Dalai Lama, X. (2010). *The universe in a single atom : the convergence of science and spirituality*. Morgan Road Books, New York.

Dambrun M., R. M. (2012). The questions of king milinda. *Journal of the American Oriental Society*, 84(93):89–102.

Davids, T. W. R. (1964). The questions of king milinda. *Journal of the American Oriental Society*, 84(4):490.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47:5–20.

Davidson, D. (1986a). A coherence theory of truth and knowledge. In LePore, E., editor, *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson*, pages 307–319. Blackwell.

Davidson, D. (1986b). A nice derangement of epitaphs. In Lepore, E., editor, *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, pages 433–446. Blackwell.

Davidson, D. (2001). *Inquiries Into Truth and Interpretation*. Clarendon Press.

Davidson, D. (2005). *Truth, Language, and History*. Collected essays. Clarendon Press.

Davidson, R. J. and Lutz, A. (2008). Buddha's brain: Neuroplasticity and meditation. *IEEE Signal Processing Magazine*, 25(1):176–174.

DeCharms, R. C. (1998). *Two views of mind : Abhidharma and brain science*. Snow Lion Publications, Ithaca, N.Y.

Durkheim, E. and Fields, K. E. (1995). *The elementary forms of religious life*. Free Press, New York.

Engle, A. B. (2009). *The inner science of buddhist practice: Vasubhandu's summary of the five heaps with commentary by Sthiramati*, volume 7. Shambhala Publications.

Ervas, F. and Ojha, O. (2019). Metaphor in argument production vs. understanding. In *Ninth Conference of the International Society for the Study of Argumentation*, pages 330–41. ISSA.

Flanagan, O. (2011). *The Bodhisattva's Brain: Buddhism Naturalized*. MIT Press, Cambridge, Massachusetts.

Fodor, J. A. (1979). *The language of thought*. Harvard University Press, Cambridge, Mass.

Fodor, J. A. (2008). *LOT 2 The Language of Thought Revisited*. Clarendon Press, Oxford.

Garfield, J. (2015a). *Engaging Buddhism: Why it Matters to Philosophy*. Oxford University Press.

Garfield, J. L. (2015b). *Engaging Buddhism : why it matters to philosophy*. Oxford University Press, New York.

Giere, R. (1999). *Science Without Laws*. Science and its conceptual foundations. University of Chicago Press.

Gould, S. J. (2007). *Rocks of ages : science and religion in the fullness of life*. [International Society for Science and Religion], [Cambridge].

Gover, T. (2018). *Problems in Argument Analysis and Evaluation*. Windsor: University of Windsor.

Hansson, S. O. (2000). Formalization in philosophy. *The Bulletin of Symbolic Logic*, 6(2):162–175.

Harrington, A. and Zajonc, A. (2008). *The Dalai Lama at MIT*. Harvard University Press, Cambridge, Mass.

Hershock, P. D. (2021). *Buddhism and Intelligent Technology*. Bloomsbury Academic, New York.

Hesse, M. (1965). Models and analogies in science. *British Journal for the Philosophy of Science*, 16(62):161–163.

Hinton, G. E. et al. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

Hofstadter, D. R. (1996). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, Inc, New York, NY, USA.

Hölzel, B. K., Carmody, J., Vangel, M., Congleton, C., Yerramsetti, S. M., Gard, T., and Lazar, S. W. (2011). Mindfulness practice leads to increases in regional brain gray matter density. *Psychiatry research*, 191(1):36–43.

Hölzel, B. K., Ott, U., Hempel, H., Hackl, A., Wolf, K., Stark, R., and Vaitl, D. (2007). Differential engagement of anterior cingulate and adjacent medial frontal cortex in adept meditators and non-meditators. *Neurosci Lett*, 421(1):16–21.

Hourihan, M. and Parkes, D. (2016). Federal R&D Budget Trends: A Summary. publisher: American Association for the Advancement of Science.

Hurley, M., Dennett, D., Adams, R., Adams, R., and Adams, A. (2011). *Inside Jokes: Using Humor to Reverse-engineer the Mind*. MIT Press.

III, S. C. W. (1986). Indeterminacy of french interpretation: Derrida and davidson. In LePore, E., editor, *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Cambridge: Blackwell.

Jacobs, S. (2010). J. B. Conant's other assistant: Science as depicted by Leonard K. Nash, including reference to Thomas Kuhn. *Perspectives on Science*, 18(3):328–351.

Jahnke, A. (2015). Who Picks Up the Tab for Science? *BU Today*. publisher: Boston University.

James, W. (1902). *Talks to teachers on psychology: and to students on some of life's ideals / By William James*. Longmans, Green, and Company London.

James, W. (1950). *The Principles of Psychology*. Number v. 1-2 in Dover Books. Dover Publications.

James, W. and Perry, R. (1912). *Essays in Radical Empiricism*. Longmans, Green, and Company.

Jeffares, B. (2010). The co-evolution of tools and minds: Cognition and material culture in the hominin lineage. *Phenomenology and the Cognitive Sciences*, 9(4):503–520.

Kabat-Zinn, J. (2011). Some reflections on the origins of mbsr, skillful means, and the trouble with maps. *Contemporary Buddhism*, 12(1):281–306.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Karunadasa, Y. (2015). *The Buddhist analysis of matter*. University of Hong Kong, Hong Kong.

Keynes, J. M. (1921). *A treatise on probability, by John Maynard Keynes*. Macmillan London.

Kiyota, M. and Jones, E. (1978). *Mahayana Buddhist meditation: theory and practice*. University Press of Hawaii.

Kuhn, T. and Wilson, K. (2001). The road since structure: Philosophical essays, 1970–1993, with an autobiographical interview. *Physics Today*, 54:53.

Kuhn, T. S. (1957). *The Copernican revolution; : planetary astronomy in the development of Western thought*. Harvard University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kuhn, T. S. (1977). *The Essential tension : selected studies in scientific tradition and change*. University of Chicago Press.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981.

Lazar, S. W., Kerr, C. E., Wasserman, R. H., Gray, J. R., Greve, D. N., Treadway, M. T., McGarvey, M., Quinn, B. T., Dusek, J. A., Benson, H., Rauch, S. L., Moore, C. I., and Fischl, B. (2005). Meditation experience is associated with increased cortical thickness. *Neuroreport*, 16:1893–7.

Lee, T. (2020). Learning to learn. *Berkeley Engineering News*.

Lewontin, R. C. (1983). The organism as the subject and object of evolution. *Scientia*, 77(18):65.

Licato, J. and Cooper, M. (2020). Assessing evidence relevance by disallowing assessment. In *Proceedings of the Ontario Society for the Study of Argumentation Conference*, volume 12, Ontario, Canada. OSSA.

Luders, E., Kurth, F., Mayer, E. A., Toga, A. W., Narr, K. L., and Gaser, C. (2012). The unique brain anatomy of meditation practitioners: alterations in cortical gyrification. *Frontiers in human neuroscience*, 6:34–34.

Ma, L. and Brakel, J. (2016). Revisiting wittgenstein on family resemblance and colour(s). *Philosophical Investigations*, 39(3):254 – 280.

Macagno, F., Walton, D., and Tindale, C. (2017). Analogical arguments: Inferential structures and defeasibility conditions. *Argumentation*, 31(2):221–243.

Mak, A. (2021). Why teslas keep striking parked firetrucks and police cars. *Slate*.

Marcus, G. and Davis, E. (2019). *Rebooting AI: building artificial intelligence we can trust*. Pantheon Books, New York, first edition edition.

Masterman, M. (1970). *The Nature of a Paradigm*, volume 4, page 59–90. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Millikan, R. G. (1984). *Language, thought, and other biological categories : new foundations for realism*. MIT Press, Cambridge, Mass.

Mitchell, M. (1993). *Analogy-Making as Perception: A Computer Model*. MIT Press, Cambridge, MA, USA.

Mitchell, M. (2020). Can gpt-3 make analogies? *Medium*.

Moleski, M. X. (2006). Polanyi vs. kuhn. . 33.2 (2006): 8-24. print. *Tradition and Discovery: the Polanyi Society Periodical*, 33(2):8–24.

Nauriyal, D. K., Drummond, M., and Lal, Y., editors (2010). *Buddhist thought and applied psychological research : transcending the boundaries*. Routledge, London.

Nishida, K. (1960). *A study of good*. Greenwood Pr., New York.

Odin, S. (1982). *Process Metaphysics and Hua-Yen Buddhism: : A Critical Study of Cumulative Penetration Vs. Interpenetration*. Suny Press.

of Scientific Research, A. F. O. (2014). Afosr 2014 technical strategic plan. *AFOSR 2014 Technical Strategic Plan*, page 1–26.

Office, A. R. (2021). Aro year in review 2020. *ARO Year in Review*.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1978). *The measurement of meaning*. University of Illinois Press, Urbana-Champaign.

Polanyi, M. (1949). Science, faith and society. *Philosophy of Science*, 16(4):353–353.

Polanyi, M. (1958). *Personal knowledge; towards a post-critical philosophy.* University of Chicago Press, Chicago.

Polanyi, M. and Sen, A. (1966). *The Tacit Dimension.* University of Chicago, Chicago, IL.

Prebish, C. S. and Kalupahana, D. (1976). Causality: The central philosophy of buddhism. *Journal of the American Oriental Society*, 96(3):463.

Purser, R. (2015). The myth of the present moment. *Mindfulness*, 6(3):680–686.

Queen, C. and King, S. (1996). *Engaged Buddhism: Buddhist Liberation Movements in Asia*. Tradition; 17; Garland Referen. State University of New York Press.

Quine, W. V. (1960). *Word and object*. Massachusetts Institute of Technology, Cambridge, Mass.

Quine, W. V. (1969). *Ontological Relativity and Other Essays*. Columbia University Press, New York.

Quine, W. V. (1973). *The roots of reference*. Open Court, LaSalle, Ill.

Quine, W. V. (1981). *Theories and things*. Harvard University Press, Cambridge, Mass.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, 60(1):20–43.

Quine, W. V. O. (1955). *Posits and Reality*, page 246. The Ways of Paradox and Other Essays. Harvard University Press, Cambridge, MA, 2nd edition.

Quine, W. V. O. (1966). *The Ways of Paradox and Other Essays*. Harvard University Press, Cambridge, MA, 2nd edition.

Quine, W. V. O. (1974). *The Roots of Reference*. Open Court Publishing Company, La Salle, Illinois.

Raichle, M. E. and Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage*, 37(4):1083–1090.

Rescher, N. (2000). *Process Philosophy: A Survey of Basic Issues*. University of Pittsburgh Press.

Ricard, M. (2009). *Quantum and the Lotus : a Journey to the Frontiers Where Science and Buddhism Meet.* Broadway.

Robertson, D. (2010). *The Philosophy of Cognitive-Behavioural Therapy (Cbt): Stoic Philosophy as Rational and Cognitive Psychotherapy*. Karnac.

Ronkin, N. (2005). *Early Buddhist Metaphysics: The Making of a Philosophical Tradition*. Routledge, New York.

Rospatt, A. v. (1995). *The Buddhist doctrine of momentariness: a survey of the origins and early phase of this doctrine up to Vasubandhu*. Number 47 in Alt- und neu-indische Studien. F. Steiner Verlag, Stuttgart.

Ruiz Fernández, J. (2018). Language as a family-resemblance concept in wittgenstein. *Philosophia (Ramat Gan)*, 47(5):1447–1455.

Russell, S. (1988). *Analogy by Similarity*, pages 251–269. Springer Netherlands, Dordrecht.

Sharp, P. B., Sutton, B. P., Paul, E. J., Sherepa, N., Hillman, C. H., Cohen, N. J., Kramer, A. F., Prakash, R. S., Heller, W., Telzer, E. H., and Barbey, A. K. (2018). Mindfulness training induces structural connectome changes in insula networks. *Scientific Reports*, 8(1):7929.

Sibert, C., Stocco, A., and Hake, H. (2021). The structured mind at rest: Evidence for the "common model of cognition" in resting state fmri. In *International Conference on Cognitive Modeling*. Society for Mathematical Psychology.

Siderits, M. (2007). *Buddhism as Philosophy : An Introduction.* Hackett, Indianapolis.

Siderits, M. (2020). Meta-cognition without a cognizer: Buddhist non-self and awareness of awareness. *Tetsugaku*, 4:103–118.

Silva, P. D. (2005). *An introduction to Buddhist psychology.* Palgrave Macmillan, Basingstoke.

Singleton, O., Hölzel, B. K., Vangel, M., Brach, N., Carmody, J., and Lazar, S. W. (2014). Change in Brainstem Gray Matter Concentration Following a Mindfulness-Based Intervention is Correlated with Improvement in Psychological Well-Being. *Front Hum Neurosci*, 8:33.

Skelac, I. and Jandric, A. (2020). Meaning as use: From wittgenstein to google's word2vec. *"Guide to Deep Learning Basics"*, pages "41–53".

Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, 9(4):465–481.

Taren, A. A., Gianaros, P. J., Greco, C. M., Lindsay, E. K., Fairgrieve, A., Brown, K. W., Rosen, R. K., Ferris, J. L., Julson, E., Marsland, A. L., Bursley, J. K., Ramsburg, J., and Creswell, J. D. (2015). Mindfulness meditation training alters stress-related amygdala

resting state functional connectivity: a randomized controlled trial. *Social Cognitive and Affective Neuroscience*, 10(12):1758–1768.

Thera, N. (1949). *Abhidhamma Studies : Buddhist explorations of consciousness and time*. Buddhist Publication Society, Kandy, Sri Lanka.

Toulmin, S. (2003a). *Return to Reason*. Harvard University Press, Massachusetts.

Toulmin, S. E. (2003b). *The Uses of Argument*. Cambridge University Press, 2 edition.

Turner, S. (2018). *Cognitive Science and the Social: A Primer*. New York, USA: Routledge.

Varela, F. J., Thompson, E., and Rosch, E. (1997). *The embodied mind : cognitive science and human experience*. MIT Press, Cambridge, Mass.

Vasubandhu, Sangpo, G., and Dhammajoti, B. (2012a). *Abhidharmakosa-Bhasya of Vasuandhu: Volume 4*. Abhidharmakosa-Bhasya of Vasubandhu. Motilal Banarsidass Publishers.

Vasubandhu, Sangpo, G., and Dhammajoti, B. (2012b). *Abhidharmakosa-Bhasya of Vasubandhu: Volume 1*. Abhidharmakosa-Bhasya of Vasubandhu. Motilal Banarsidass Publishers.

Vasubandhu, Sangpo, G., and Dhammajoti, B. (2012c). *Abhidharmakosa-Bhasya of Vasubandhu: Volume 2*. Abhidharmakosa-Bhasya of Vasubandhu. Motilal Banarsidass Publishers.

Vasubandhu, Sangpo, G., and Dhammajoti, B. (2012d). *Abhidharmakosa-Bhasya of Vasubandhu: Volume 3*. Abhidharmakosa-Bhasya of Vasubandhu. Motilal Banarsidass Publishers.

Wallace, B. (2009a). *Meditations of a buddhist skeptic : a manifesto for the mind sciences and contemplative practice.* Columbia University Press, New York.

Wallace, B. (2010). *Hidden dimensions : the unification of physics and consciousness*. Columbia University Press, New York; Chichester.

Wallace, B. (2013). *Meditations of a Buddhist Skeptic: A Manifesto for the Mind Sciences and Contemplative Practice*. Columbia University Press.

Wallace, B. A. (2003). *Choosing reality : a Buddhist view of physics and the mind*. Snow Lion Publications, Ithaca, N.Y.

Wallace, B. A. (2004). *The Taboo of Subjectivity : Toward a New Science of Consciousness*. Oxford University Press, USA, Oxford.

Wallace, B. A. (2007). *Buddhism & Science : Breaking New Ground*. [International Society for Science and Religion], [Cambridge].

Wallace, B. A. (2009b). *Mind in the Balance: Meditation in Science, Buddhism, and Christianity*. Columbia University, New York.

Wallace, B. A. and Hodel, B. (2009a). *Contemplative science : where Buddhism and neuroscience converge*. Columbia University Press, New York; Chichester.

Wallace, B. A. and Hodel, B. (2009b). *Embracing mind : the common ground of science and spirituality*. Shambhala ; Publishers Group UK [distributor], Boston, Mass.; Enfield.

Walton, D. (2014). *Argumentation Schemes for Argument from Analogy*, pages 23–40. Springer.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

Wilson, E. O. (1998). The biological basis of morality. *The Atlantic Monthly*, pages 53–70.

Wittgenstein, L. and Anscombe, G. E. M. (2000). *Philosophical investigations: the English text of the third edition*. Prentice Hall, Englewood Cliffs, N.J, 3. ed edition.

XIV, D. L., Houshmand, Z., Livingston, R. B., and Wallace, B. A. (1999). *Consciousness at the crossroads : conversations with the Dalai Lama on brain science and Buddhism*. Snow Lion Publications, Ithaca, N.Y.

Zeidan, F., Grant, J. A., Brown, C. A., McHaffie, J. G., and Coghill, R. C. (2012). Mindfulness meditation-related pain relief: evidence for unique brain mechanisms in the regulation of pain. *Neuroscience letters*, 520(2):165–173.

Zhou, Y. and Yurovsky, D. (2021). A common framework for quantifying the learnability of nouns and verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43. Cognitive Science Society.