Media Engineering and Technology Faculty
German University in Cairo

Algebraic Logic of Desire

Bachelor Thesis

Author: Ahmed Haitham Ibrahim
Supervisors: Prof. Haytham Osman Ismail

Submission Date: 19 May, 2024

Media Engineering and Technology Faculty
German University in Cairo

Algebraic Logic of Desire

Bachelor Thesis

Author: Ahmed Haitham Ibrahim
Supervisors: Prof. Haytham Osman Ismail

Submission Date: 19 May, 2024

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

---

Ahmed Haitham Ibrahim
19 May, 2024

Acknowledgments

As I reflect on the journey that has brought me to this moment, I am filled with gratitude for the people who have supported me along the way. First and foremost, I want to express my deepest thanks to my supervisor, Professor Haythem Ismail. His guidance, wisdom, and patience have been a constant source of comfort and inspiration. His ability to challenge me to push beyond my limits has been invaluable, and I am so grateful to have had the opportunity to learn from him.

I also want to thank my family and friends, who have been my rock throughout this process. Their love, encouragement, and belief in me have given me the strength and motivation to keep going, even when the road ahead seemed uncertain, even when I doubted myself. This thesis is dedicated to you.

I am also deeply grateful to the researchers who have come before me, whose work has laid the foundation for my own research. Their contributions have inspired me and given me the information needed to pursue my own ideas and contribute to my own piece of the puzzle.

Finally, I want to acknowledge the countless hours, late nights, and early mornings that have gone into this thesis. It has been a labor of love, and I am proud of what I have accomplished. I hope that my work will make a meaningful contribution to the field and inspire others to pursue their own research passions.

Abstract

Desire is a fundamental aspect of human motivation and agency, yet its complexities have proven challenging to capture in formal representations. This thesis addresses the study of desire using the algebraic language Log_AC_n.

Building upon previous work, we propose a formal framework that enables a more comprehensive representation and analysis of desire's properties. Our approach provides a rigorous and expressive language for describing desire, with potential implications for both philosophical inquiry into motivation and agency, and the development of more sophisticated and human-like intelligent agents in artificial intelligence research. This thesis presents a systematic exploration of desire within Log_AC_n framework, proving key propositions about its properties and behavior.

Contents

Chapter 1

Introduction

1.1 Motivation

The concept of desire, a fundamental aspect of human motivation, has been explored through a multitude of philosophical and psychological lenses. Action-based, learning-based, pleasure-based, good-based, and attention-based theories have all added to the understanding of desire, yet historically a unified understanding remains elusive.

In parallel, various logics for belief and intention have been proposed whereas logics of desire had less attention in artificial intelligence literature. The field of artificial intelligence (AI) has always been interested in representing and reasoning about motivational attitudes in intelligent agents. While traditional AI planning systems and Belief-Desire-Intention (BDI) models have incorporated the notion of desire, these approaches often lack the nuanced explanation required to capture the full spectrum of human desire.

Beyond theoretical exploration, understanding desire holds significant practical implications, particularly in the realm of artificial intelligence. One motivation behind this research is to deepen our understanding of desire itself, recognizing its complexity and centrality to human behavior. If AI agents are to interact with humans, a detailed grasp of desire within their knowledge base could lead to more sophisticated, human-like intelligence. Even if AI agents cannot directly experience desire, comprehending the concept allows them to understand human motivations and decision-making processes. For instance, an AI system capable of explaining its actions in terms of desire ("I chose this path because I desired the most efficient outcome") would bridge the communication gap between humans and machines.

1.2 Aim

This thesis aims to establish a more comprehensive framework for understanding and analyzing desire. We will achieve this by extending the study of desire using the algebraic

language Log_AC_n. The thesis will begin with a thorough review of existing literature on different logics of desire, examining various approaches and highlighting their strengths and limitations. Building upon this foundation, we will then delve into the development of an algebraic logic of desire within the Log_AC_n framework. This approach will allow us to formally represent the complexity of desire.

1.3 Organization of the Thesis

The thesis is organized as follows. In Chapter 2 the Background, we will delve into what is desire, and what well-known theories of desire are some of them are formal and others are not. In Chapter 3, We will try to present the language Log_AC_n and illustrate why the language could simplify many concepts when it comes to desire formality then we will move to prove some propositions about desire using the notion of desire presented in the language.

Chapter 2

Theories of Desire

The context of desire necessitates an exploration of diverse theories and frameworks. We will review together some of the classical theories of desire. This section aims to traverse the field of desire theories. We lay the groundwork for a detailed algebraic logic of desire and future possible applications. The term "dispose" throughout the section refers to the inclination or tendency to act in a certain way to fulfill a desire.

2.1 Pleasure-Based Theories of Desire

As mentioned above, philosophers who complain that action-based theories of desire fail to distinguish judgments of good from desire sometimes suggest that pleasure may play a role in such distinction.
In its simplistic form, Pleasure-based theories of desire propose that desires are fundamentally linked to the experience of pleasure and displeasure.

For an organism to desire p is for the organism to be disposed to take pleasure in it seeming that p and displeasure in it seeming that $not - p$.

According to these theories, a person moved by a desire always takes pleasure in the desired outcome or eagerly anticipates its satisfaction, whereas a person moved only by a judgment of goodness does not necessarily share these feelings.

Carolyn Morillo [10] presents a unique perspective on desire, particularly through a pleasure-based theory. This theory departs from traditional views by arguing that desires are independent of action and are intimately tied to experiences of pleasure. Morillo's argument begins with conceptual considerations, suggesting that desires should be viewed as non-trivial explainers of action. In other words, desires are not merely outcomes of actions but play a fundamental role in motivating and explaining why we act in particular ways.

Morillo draws on neuroscientific evidence, specifically focusing on the release of dopamine by the brain's reward system during episodes of pleasure. Dopamine is a neurotransmitter associated with feelings of pleasure and reward, and its release is linked to various pleasurable experiences, such as eating delicious food or engaging in enjoyable activities.

Morillo contends that these neural events associated with pleasure, particularly the release of dopamine, are not only correlated with desires but also serve as their causal origin when combined with belief. Belief here refers to cognitive processes and mental representations about the likelihood of achieving desired outcomes.

By linking episodes of pleasure to neural events and their role in motivating action, Morillo proposes that desires can be understood as episodes of pleasure. In other words, when we experience pleasure, particularly through the release of dopamine, it directly influences our desires and motivates us to take action to pursue those pleasurable experiences.

This perspective challenges traditional views of desires and provides a neuroscientific basis for understanding their underlying mechanisms. Pleasure-based theories of desire face a challenge since some philosophers believe that the net increase in satisfaction of desire is the standard cause of pleasure, and the pleasure caused perhaps represents this change in desire satisfaction. If such views are correct then it must be the case that desire is distinct from pleasure just like causes are distinct from their effects.

2.2 Learning-Based Theories of Desire

Learning-based theories of desire propose that desires are linked to reward-based learning mechanisms in the brain. According to Timothy Schroeder [14], desires involve representations used to drive reward-based learning, mediated by the dopamine-releasing reward system.

For an organism to desire p is for it to use representations of p to drive reward-based learning.

These theories suggest that desires can exist independently of typical associated behaviors or feelings, focusing on the processes involved in learning.

Critics of learning-based theories suggest that accepting these theories might mean we have to let go of pre-established limits on how we think about desires. It could also mean accepting that desires might exist even if they don't always come with the usual features we associate with them.
Learning-based theories make sense if we agree that desires are a natural thing that can show up in different ways and link to actions and feelings. But for some people, this idea doesn't match up with what they think desire is supposed to be.

2.3 Action-Based Theories of Desire

The simple theory of desire is that desire is what makes you act and according to that the most important thing in desire is the tendency to act or to have a disposition to act. If a person wants to get a cake, then a cake is what the person desires however a person can desire a cake without even trying to get it, He might be craving cake while he is not willing to go to the bakery. This leads to a simple action-based theory of desire

For an organism to desire p is for the organism to be disposed to act so as to bring about p.

This theory has been criticized for not being restrictive enough, For example, if a woman has a predisposition to stutter, the argument above implies that she wants to stutter simply because she is predisposed to do so. For these reasons, many people prefer a more complex version of an action-based explanation of desire.

For an organism to desire p is for the organism to be disposed to take whatever actions it believes are likely to bring about p.

Many philosophers suggested that desires are only a psychological state that can initiate an action. So identifying desires as psychological states disposing us to actions might be wrong or not accurate. Some of these philosophers have focused on 'true desires' which refer to desires that align with our genuine, deeply held values, and preferences. However, according to this perspective, even these "true desires" do not encompass all possible motivational states.

Frankfurt's 1971 article delves into a tricky philosophical problem about desires and actions. He talks about situations where people do things based on desires they may not fully agree with. Even though they act on these desires, Frankfurt suggests that it doesn't always mean they have true free will. The big question is how people can both accept and reject desires without needing to believe in a separate part of their mind making choices. This makes us think more about how people make decisions and control their own actions [1].

There may be other factors, such as habit, or situational pressures, that can also drive our actions independently of our desires. While desires typically involve wanting or craving something, judgments of goodness are related to evaluating actions, objects, or outcomes as morally or ethically desirable which we will see in the good-based theories of desire.

But according to these philosophers, desires, and judgments of goodness are two different things. Hence action-based theories of desire fail to make such an important distinction, to overcome the issues an action-based theorist may say that belief in goodness cannot move people to act or there's incoherence in the principles by which one can decide what's good or what's not or one in the principles that dispose us to act.

2.3.1 Frankfurt's Hierarchy and Ownership

Frankfurt's model suggests that desires become truly our own when we want to act on them, which he calls "second-order volitions" However, just having these higher-order desires isn't enough to claim ownership of our desires. We need to explain why some of these higher-order desires have the power to guide our actions without causing a problem of never-ending questioning.

a challenge arises regarding whether evaluative judgments or higher-order desires should be the primary basis for claiming ownership or rejection of desires. To address this, some suggest separating Frankfurtian higher-order attitudes from evaluative judgments. Yet, it is argued that maintaining some link between higher-order desires and evaluative judgments is vital for effective functioning, even if not universally required for authority or desire ownership.

Donald Davidson [2] proposed that desiring involves an "all-out" evaluative judgment, interpreted as judging one option as strictly better than its alternatives. However, this strict evaluative judgment faces a buridan problem in cases where options are seen as equally good.

Frankfurtian commitments which involve rejecting a desire based on higher-order attitudes, can address such uncertainties and move beyond prior value judgments. These commitments are expected to be stable and play appropriate roles in an agent's practical thought and action over time. Reflective arrival at these commitments may involve considering a wide range of thoughts, feelings, and inclinations, rather than simply brute picking.

2.4 Good-Based Theories of Desire

While some philosophers hold that desires need to be separated from the judgment of goodness. Others propose that desires and judgments of goodness are closely tied. One of them is Socrates, he said to want something is simply to think it's good, and based on that it's simple to say

$$\text{For an organism to desire } p \text{ is for it to believe } p \text{ is good.}$$

Advocates of such theories argue that we are motivated to pursue what we judge to be good simply because we judge it to be good.

2.4.1 Anti-Humean's challenge

Lewis [6] has tackled a challenge of such theories the anti-Humean challenge presents a scenario where a decision is made against a warm desire (e.g., the desire for an old friend

to get a job) in favor of a cold desire (e.g., the desire to hire the best candidate). The Humean response argues that both warm and cold desires can motivate action, with the latter often outweighing the former. However, the anti-Humean contends that beliefs about what is good necessarily entail corresponding desires, either through identity or necessary connection.

The Anti-Humean's position is challenged in two ways. First, it is suggested that beliefs about what is good may not qualify as true beliefs but rather as desires by courtesy, functioning differently from genuine beliefs. Alternatively, it is proposed that while beliefs about what is good may be genuine beliefs, their connection to desires could be contingent rather than necessary, with explanations provided for why they often coexist. The debate remains unresolved.

2.4.2 Desire-as-Belief and Desire-as-Conditional-Belief

Decision Theory provides a formal framework for understanding belief, desire, and decision-making. It offers mathematical and logical tools to analyze how agents make choices based on their beliefs and desires.[7]

The argument against the Desire-as-Belief Thesis, when applied to Decision Theory, presents a collision that challenges the compatibility of the two concepts(Decision theory and Desire as belief).

The collision arises when Decision Theory is applied to a hypothetical agent named Frederic, who is assumed to be entirely motivated by beliefs about what is good.

In the case of Frederic, his beliefs about what is good directly determine his desires. However, Decision Theory demonstrates that under certain conditions, Frederic cannot simultaneously change both his opinion (about whether a proposition is true) and his desire (about whether a proposition should be true). This contradicts the Desire-as-Belief Thesis, which asserts a necessary connection between beliefs and desires.

Challenges for good-based theories of desire include explaining the relationship of desires to non-humans. On one hand, it would seem that rats desire to get away from cats, desire to be around other rats, and the like. On the other hand, it would seem that rats do not represent anything as good (they would both seem to lack the concept of goodness and lack a perceptual-style representation of goodness that would be well poised to generate such a concept).

But if rats can desire without representing the good, then why would people be different? The exploration of solutions to such puzzles is ongoing in the philosophical literature.

2.5 Attention-Based Theories of Desire

Another evaluatively loaded theory of desire has been proposed by T.M. Scanlon. Dubbed a theory of desire in the 'directed-attention' sense, this theory links desires to reasons,

rather than goodness. But the theory does so through its characterization of how desire plays its most important role, which is its role in directing the attention of the subject who desires.

For an organism to desire p is for the thought of p to keep occurring to the organism in a favorable light so that its attention is directed insistently toward considerations that present themselves as counting in favor of p.

In Scanlon's view [12], reasons are factors that weigh in favor of propositions, it follows from this theory that a desire p exists if one's attention is directed insistently toward apparent reasons to have it be the case that p. This is where the evaluative element enters the theory.
Technically, Scanlon does not present a complete attention-based theory of desire, but only an attention-based sufficient condition for the existence of a desire. Perhaps this is because Scanlon sees his theory as best suited to characterizing desires that are present. A theory of standing desires that follows Scanlon's lead might look like this

For an organism to desire p is for it to be disposed to keep having its attention drawn to reasons to have p, or to reasons to avoid $not - p$.

One of the theory's flaws may be its emphasis on a single type of attention. According to the theory, desire has the characteristic effect of directing attention towards reasons to satisfy the desire. But desire has an impact on other types of attention as well.

2.6 Preferences and Desires

A fundamental conflict arises regarding the primacy of desires or preferences. While decision theorists typically regard pairwise preferences as fundamental, others argue for the foundational nature of desires. This discrepancy impacts the derivation of preferences from desires and vice versa.[13]

Some proponents assert that preferences are basic due to their ease of introspection and immediate influence on behavior. Unlike desires, which exhibit varying strengths, preferences are thought to be grounded in immediate sensations of preference between objects, as suggested by Von Neumann and Morgenstern.

Despite assertions about the basic nature of preferences, standard decision theory suggests that deriving preferences from pairwise comparisons necessitates a vast number of basic pairwise preferences. This complexity has led some, like John Pollock, to advocate for the psychological realism of basic desires.

Research shows that preferences can change based on the situation, suggesting that preferences may not be as stable as previously thought. For example, how choices are presented can greatly influence preferences, leading researchers to question how stable preferences really are when the context changes.

Chapter 3

Formal Theories of Desire

3.1 Desirable Propositions

The classical logic viewpoint defines desirable propositions as those implying the agent's end goal, tethering the desire to complete knowledge. This 'All-or-Nothing Approach' suggests that if an action cannot guarantee full goal achievement, it cannot be considered desirable. However, this perspective presents limitations for real-world agents, whether they are humans or AI systems, we hardly ever know everything perfectly. So, we end up being too careful or not doing anything at all because we're not sure.

Previous research in qualitative decision theory has primarily concentrated on comparing the preferences among different propositions. Preference semantics allow for determining which proposition is more desirable than another, facilitating decision-making accordingly. However, they argue that desirability holds equal importance to preference in rational decision-making. There exist significant distinctions between desirable and undesirable propositions.

Qualitative decision theory can represent this contrast, it fails to explain why one proposition is desirable while the other is not; it only indicates which proposition is more preferable based on our preferences across different scenarios.

Zhou and Chen [15] refined this classical approach, introducing a more detailed framework to address its limitations and showcase the relationship between the agent's goal and desirable propositions. Using what they have introduced by partial implication they were able to redefine desirable propositions to address the limitation of incomplete knowledge. It allows propositions that advance an agent toward a goal having incomplete knowledge and having uncertainty about total success.

3.1.1 Formal syntax

They have limited their discussion within a propositional language L. The formulas of L are formed from sets of atoms, Atom=$\{x_1, x_2, \dots\}$, and logical connectives $\neg$,$\vee$, $\wedge$,$\Rightarrow$, and $\Leftrightarrow$.

They defined [15] A as a subset of atoms, and with that subset of atoms they have created literals from A $L(A) = A \cup \{ \neg x \mid x \in A\}$, simply the atoms themselves and their negations, so for each atom x in A, we add both (x) and its negation ($\neg$ x).

The set of formulas Γ, Γ' represent sets of formulas in the language L. These sets contain formulas that don't contradict each other.

If we have a formula P or a set of formulas Γ. **Atom**(P) tells us which atoms appear in P, and **Atom**(Γ) tells us which atoms appear in the set Γ.

They have set two important definitions Γ implicant and the Γ-prime implicant Γ implicant is an implicant of formula P for set Γ is a consistent conjunction of literals of π if the conjunction π is consistent with the set Γ and When combined with the set Γ, the conjunction π satisfies the formula P.

We denote the set of implicants of P to Γ as Γ[P]. An implicant is a partial truth-value assignment, meaning it may not assign a truth value to every atom. This reflects incomplete knowledge or recognition by an agent. However, sometimes complete knowledge isn't needed, and a formula's truth value can be determined even with an implicant. For instance, if Γ is empty, an implicant of the formula $(x \vee y)$ could be {x}, where only x is assigned a truth value. In this case, the truth value of $(x \vee y)$ doesn't depend on y. Thus, an implicant can be seen as a simplified representation of the relevant state.

Γ-prime implicant is A prime implicant of formula P with respect to set Γ is a consistent conjunction of literals π if π is an implicant of P with respect to Γ and there isn't another implicant π' of P under Γ such that π' is a proper subset of π.

We denote the set of prime implicants of P with respect to Γ as Γ(P). It is important to note that every prime implicant is also an implicant, so Γ(P) is a subset of Γ[P]. If the set of atoms in P combined with the set of atoms in Γ is finite, then Γ(P) is also finite.

In simple terms, a prime implicant of P represents an exact way to achieve P, where every atom in the prime implicant is necessary. Each atom in a prime implicant is essential to achieving P. If another proposition makes some literals in a prime implicant true without making the rest false, it is considered useful to P and harmless otherwise. A subset of a prime implicant is called a "piece" of P, which is simpler than the original goal P. If Γ is empty, π is simply called a prime implicant of P.

Let - π be the set of negations of the literals in π. Similar to Lmp4c, we define partial implication based on prime implicant. We say that P partially implies Q under set Γ, denoted by $\Gamma \models P \succ Q$, if Γ(P) is not empty, meaning there are prime implicants of P under Γ and For each prime implicant π of P in Γ(P), there exists a prime implicant π' of Q in Γ(Q) such that π and π' have some common atoms ($\pi \cap \pi'$ is not empty), and there are no common negated atoms ($\pi \cap$ -π' is empty).

The prime implicant π of P serves a dual purpose. Firstly, it represents all possible situations consistent with Γ in which P is true. Secondly, it indicates that P is both useful (it contributes to Q) and harmless (it doesn't contradict Q) to Q through its intersection with π'. Thus, P is useful and harmless to Q through π and π'.

3.1.2 Partial Implication Relationships

Theorem 1 (Equivalence) *If* $\Gamma \models P \leftrightarrow Q$, *then* $\Gamma \models P \succ R$ *implies* $\Gamma \models Q \succ R$; $\Gamma \models R \succ P$ *implies* $\Gamma \models R \succ Q$.

two propositions P and Q have the same truth value under the conditions specified by set Γ, *then any proposition R that is partially implied by P is also partially implied by Q, and vice versa.*

Theorem 2 establishes a crucial relationship between logically equivalent propositions and their partial implications. This result highlights the connection, of propositions within a given set of conditions, and shows the consistency and coherence of logical reasoning under different scenarios.

Theorem 2 (Relationship to classical implication) *If* $\Gamma \models P \Rightarrow Q$ *and both P and Q are non-trivial under* Γ, *then* $\Gamma \models P \succ Q$.

From theorem 3 it follows that partial implication is an extension of classical implication.

Theorem 3 (Non-triviality) *If P or Q is trivial under* Γ, *then that* $\Gamma \models P \succ Q$ *doesn't hold.*

This theorem illustrates that every trivial formula does not partially imply other formulas and can not be partially implied either.

Theorem 4 (Disjunctive decomposition) *If* $\boldsymbol{Atom}(P) \cap \boldsymbol{Atom}(Q)$ *is empty, and P, Q are non-trivial, then* $\models P \succ (P \vee Q)$.

Theorem 5 (Conjunctive decomposition) *If* $\boldsymbol{Atom}(P) \cap \boldsymbol{Atom}(Q)$ *is empty, and P, Q are non-trivial, then* $\models P \succ (P \wedge Q)$.

Lemma 1 *Suppose* Γ *is empty, if* $\boldsymbol{Atom}(P) \cap \boldsymbol{Atom}(Q)$ *is empty, then* $\Gamma\ (P \vee Q) = \Gamma\ (P) \cup \Gamma\ (Q)$, $\Gamma\ (P \wedge Q) = \{\pi \cup \pi' \mid \pi \in \Gamma\ (P), \pi\ ' \in \Gamma\ (Q)\}$.

Proposition 1

(p-1) $\models\ x \Rightarrow (x \wedge y) \vee (\neg x \wedge \neg y)$;

(p-2) $\models\ \neg x \Rightarrow (x \wedge y) \vee (\neg x \wedge \neg y)$;

Shows that a condition and its negation can partially imply the same proposition.

Proposition 2 (Relevance) *If* $\models P \succ Q$, *then* $\boldsymbol{Atom}(P) \cap \boldsymbol{Atom}(Q)$ *is not empty.*

Unlike relevance logic, partial implication focuses on "partial" relationships between formulas. The antecedent needn't imply the consequent. For example, (p-1) and (p-7) show that some partial implication relationships don't hold in relevance logic.

(p-2) $\models x \succ (x \wedge y)$;

(p-7) $\models (x \wedge z) \succ (x \wedge y)$;

Proposition 3 (Non-monotonic) *There exists a formula P and a* Γ *-implicant* π *of P such that* π *is not a* Γ *'-implicant of P, where* $\Gamma' = \Gamma \cup \{Q\}$.

Monotonicity typically means that adding additional information (in this case, proposition Q to set Γ) should only strengthen or maintain the existing relationships. However, Property 9 demonstrates a case where adding proposition Q weakens or alters the relationship between formula P and its implicants. An example provided in the property showcases this scenario: when Γ is empty, P = x $\wedge$ y, Q = x, and $\pi = \{x, y\}$. Here, π is an implicant of P with respect to Γ, but when Q is added to Γ to form Γ', π is no longer an implicant of P under Γ'. This non-monotonic behavior extends to prime implicants as well, indicating that it is inherent within partial implication semantics. While non-monotonicity is often associated with beliefs, in the context of desirable propositions, it reflects the dynamic nature of how additional information can affect the relationship between propositions and their applicants.

Proposition 4 (Decomposition) *If any intersection of sets* $\boldsymbol{Atom}(P_1)$, $\boldsymbol{Atom}(P_2)$ ... $\boldsymbol{Atom}(P_n)$ *is empty, we can replace these formulas with atoms.*

if the atoms involved in several formulas do not overlap at all, they can be treated independently, and the formulas can be simplified by replacing them with the individual atoms. This reduction in complexity reflects the independence of knowledge held by the agent regarding different aspects represented by these formulas. This property allows for streamlining partial implication semantics by breaking down complex formulas into simpler components, thereby facilitating easier analysis and interpretation.

Proposition 5 (Non-transitivity) $\Gamma \models P \succ Q$ *and* $\Gamma \models Q \succ R$ *doesn't imply* $\Gamma \models P \succ R$.

even if P contributes to Q and Q contributes to R, it doesn't guarantee that P contributes to R in the same way. This lack of transitivity arises because the usefulness of P with respect to Q may not necessarily translate to the usefulness of Q with respect to R in certain situations. An example provided in the property illustrates this concept: when Γ is empty, both $\models x \succ (x \wedge y)$ and $\models (x \wedge y) \succ y$, hold true, but $\models x \succ y$, does not hold. This demonstrates how the partial implication relationship between P and Q, and between Q and R, does not imply the partial implication relationship between P and R.

Definition 1 (Strict desirable propositions) *Let Q be the agent's desire and Γ is the agent's belief set. If $\Gamma \models P \rightarrow Q$, then the P is called a strict desirable proposition of the agent Q.*

Definition 2 (Partial desirable propositions) *Let Q be the agent's desire and Γ be the agent's belief set. If $\Gamma \models P \succ Q$, then the P is called a partial desirable proposition of the agent Q.*

Theorem 4 showed that all nontrivial cases of strict desirable propositions are included in partial desirable propositions. However, strict desire propositions have limitations as they don't cover all interesting cases, Consider the example of Alice's goal to have milk and bread, but both are unavailable at home. While Alice knows there's bread at the store, she's uncertain about the availability of milk. In classical planning, Alice might refrain from action due to incomplete knowledge. However, under the criterion of partial desirable propositions, "going to the store" becomes a rational alternative because obtaining bread partially fulfills Alice's goal.

3.2 BDI architecture

BDI architectures are systems that prioritize beliefs (B), desires (D), and intentions (I) in the design of rational agents. Bratman's theory emphasizes the significance of intentions in practical reasoning, considering them as partial plans of action committed to by the agent to achieve their goals. Desire, intentions, and beliefs are distinct psychological constructs, but they are closely related and often influence each other.

1. **Desire** refers to a strong feeling of wanting to have or achieve something. It is a motivational force that drives individuals to pursue certain goals or outcomes. Desires can be immediate or long-term, and they can be influenced by various factors such as emotions, personal values, and external influences.

2. **Intentions** are a person's plans or commitments to perform a particular action in the future. They represent a conscious decision to act in a certain way to achieve a desired outcome. Intentions are often based on desires and are shaped by beliefs about the feasibility and consequences of the intended actions.

3. **Beliefs** are cognitive representations of how individuals understand the world around them. They encompass thoughts, attitudes, and perceptions about oneself, others, and the environment. Beliefs can be factual. Beliefs influence both desires and intentions by shaping what individuals perceive as desirable and feasible.

Cohen and Levesque formalized aspects of Bratman's theory [11], defining intentions in terms of temporal sequences of an agent's beliefs and goals. They distinguish between

fanatical commitment, where goals are maintained until achieved or deemed unachievable and relativized commitment, where goals may be dropped under specified conditions. Three crucial elements are highlighted: treating intentions as first-class citizens alongside beliefs and goals, distinguishing between the agent's choice of actions and the environment's determination of outcomes, and specifying an interrelationship between beliefs, goals, and intentions to avoid commitment to unwanted side effects.

the paper presents a temporal [11] structure called a time tree to model the world, with branches representing choices available to the agent at each moment. Events transform one time point into another, where primitive events are directly performable by the agent, and non-primitive events allow for the partial nature of plans and hierarchical plan development. Modal operators similar to Computation Tree Logic (CTL) are introduced to describe structures representing agent choices and actions over time. Two types of formulas are distinguished: state formulas and path formulas. The paper introduces two modal operators, "optional" and "inevitable," which operate on path formulas. A path formula ϕ is considered optional if, at a particular time point in a time tree, ϕ holds true for at least one path emanating from that point. Conversely, ϕ is deemed inevitable if it holds true for all paths emanating from that point. standard temporal operators such as $\bigcirc$ (next),$\Diamond$(eventually), $\Box$ (always), and $\cup$ (until), which operate over both state and path formulas, providing a formalism for reasoning about temporal properties and relationships within the BDI architecture.

Beliefs, goals, and intentions are modeled within this framework. Belief-accessible worlds represent what the agent believes to be possible, while goal-accessible worlds represent the agent's consistent goals and chosen desires that are believed to be achievable. Intentions are represented by sets of intention-accessible worlds, indicating the agent's committed attempts to realize certain courses of action, which must be compatible with her goals. The paper adopts a notion of strong realism, where the agent believes she can optionally achieve her goals through careful choices of events. Compatibility between belief-accessible, goal-accessible, and intention-accessible worlds is enforced to ensure coherence in the agent's decision-making process. Different schemes for forming, maintaining, and revising beliefs, goals, and intentions are discussed, which determine the behavioral characteristics of different types of agents.

3.2.1 Formal syntax

the paper [11] extends the propositional branching-time logic CTL(see more [8]) to reason about programs, first by introducing a first-order variant and then by extending it to a possible-worlds framework with modal operators for beliefs, goals, and intentions. While the completeness of this extension is not addressed, the main focus is on providing expressive semantics for intentions and investigating certain axioms relating intentions to beliefs and goals within this framework.

The logic introduced in this framework consists of two types of formulas: state formulas, which are evaluated at a specific time point in a given world, and path formulas, which

are evaluated along a specific path in a given world. State formulas include first-order formulas, combinations of state formulas with logical connectives, event-type formulas, and modal operators for beliefs, goals, and intentions. Path formulas include any state formula and combinations of path formulas with temporal operators such as negation, disjunction, existential quantification, and modal operators like optional. A state formula can be defined as follows:

- Any first-order formula is a state formula.
- If ϕ_1 and ϕ_2 are state formulas and x is an individual or event variable, then $\neg\phi_1$, $\phi_1 \vee \phi_2$, and $\exists x\phi_1$ are state formulas.
- If e is an event type then succeeds(e), fails(e), does(e), succeeded(e), failed(e), and done(e) are state formulas;
- If ϕ is a state formula then **BEL**(ϕ), **GOAL**(ϕ), and **INTEND**(ϕ) are state formulas.
- If ψ is a path formula, then optional(ψ) is a state formula.

A path formula can be defined as follows:

- Any state formula is also a path formula.
- If ψ_1 and ψ_2 are path formulas, then $\neg\psi_1$, $\psi_1 \vee \psi_2$, $\psi_1 \cup \psi_2$, $\Diamond\psi_1$, $\bigcirc\psi_1$ are path formulas.

The formulas "succeeded(e)" and "failed(e)" indicate whether event "e" was successful or unsuccessful in the immediate past. "done(e)" signifies whether event "e" occurred in the immediate past, regardless of its success. Similarly, "succeeds(e)", "fails(e)", and "does(e)" refer to whether event "e" is expected to succeed, fail, or occur in the immediate future. The operators **BEL**, **GOAL**, and **INTEND** represent the beliefs, goals, and intentions of the agent, respectively.

3.2.2 Semantics of World

An interpretation is represented as a tuple $M = (W, E, T, \succ, U, B, G, I, \Phi)$, where:

- W is the set of worlds.
- E is the set of primitive event types.
- T is the set of time points.
- $\succ$ is a binary relation on time points.

- U is the universe of discourse.
- Φ Mapping of first-order entities to elements in U for any given world and time point.
- B, G, and I are relations mapping the agent's current situation to her belief, goal, and intention-accessible worlds, respectively.

Each world w in W is called a time tree and is represented as $< T_w, A_w, S_w, F_w >$, where:

- T_w is a set of time points in the world w.
- A_w is the same as $\succ$ but restricted to time points in T_w.
- A full path in a world w is an infinite sequence of time points such that adjacent time points belong to A_w.
- The functions S_w and F_w map adjacent time points to events in E_w, the value of S_w represents the event that successfully occurred between those time points. Similarly, F_w represents the failure of the event between those time points.

A sub-world is defined to be a sub-tree of a world with the same truth-assignment of formulas. A world w' is a sub-world of the world w, denoted by $w' \subseteq w$, if and only if:

1. $T_{w'} \subseteq T_w$: The set of time points in w' is a subset of the set of time points in w.
2. For all $u_w \in T_{w'}$, $\Phi(q, w', u) = \Phi(q, w, u)$, where q is a predicate symbol.
3. For all $u \in T_{w'}$, $R_u^w = R_u^{w'}$, where R refers to any of B, G, I relations.
4. $A_{w'}$ is A_w restricted to time points in $T_{w'}$, and similarly for S_w and F_w.

Consider an interpretation M with a variable assignment v. The semantics of first-order formulas are defined as follows:

- $M, v, w_t \models (y_1, ..., y_n)$ if $(v(y_1), ..., v(y_n))$ satisfies the predicate formula q at time t.
- $M, v, w_t \models \neg\phi$ if $M, v, w_t \not\models \phi$.
- $M, v, w_t \models \phi_1 \vee \phi_2$ if $M, v, w_t \models \phi_1$ or $M, v, w_t \models \phi_2$.
- $M, v, w_t \models \exists_i \phi$ if there exists some d such that M, v extended with i mapped to d satisfies ϕ.
- $M, v, (w_{t_0}, w_{t_1}, \dots) \models \phi$ if $M, v, w_{t_0} \models \phi$.
- $M, v, (w_{t_0}, w_{t_1}, \dots) \models \bigcirc\psi$ if $M, v, (w_{t_1}, \dots) \models \psi$.

- $M, v, (w_{t_0}, w_{t_1}, \dots) \Diamond \psi$ if there exists $k \geq 0$ such that $M, v, (w_{t_k}, \dots) \models \psi$.
- $M, v, w_{t0} \models (\psi_1 \cup \psi_2)$ if either ψ_2 holds from some point onward and before that, ψ_1 holds, or ψ_1 holds indefinitely.
- $M, v, w_{t_0} \models \text{optional}(\psi)$ if there exists a full path $(w_{t0}, w_{t1}, \dots)$ such that $M, v, (w_{t_0}, w_{t_1}, \dots) \models \psi$.

The formula inevitable(ψ) is defined as $\neg\text{optional}(\neg\psi)$, and $\Box\psi$ is defined as $\neg\Diamond\neg\psi$.

Formulas with no positive occurrences of inevitable (or negative occurrences of optional) outside the scope of modal operators **BEL**, **GOAL**, or **INTEND** are termed O-formulas (α). Conversely, I-formulas (β) contain no positive occurrences of optional.

3.2.3 Semantics of Events

Event types represent transformations from one time point to another, known as the dynamics of a system. The successful or failed execution of events is denoted by formulas such as "succeeded(e)" and "failed(e)". Notably, the absence of an event is distinct from its failure. Failures can significantly impact the system, potentially requiring the agent to replan. For instance, the consequences of successfully robbing a bank differ from failing to rob it or not attempting to rob it at all.

- $M, v, w_t \models \text{succeeded}(e)$ if there exists t_0 such that $S_w(t_0, t_1) = e$.
- $M, v, w_t \models \text{failed}(e)$ if there exists t_0 such that $F_w(t_0, t_1) = e$.
- $M, v, w_t \models \text{done}(e)$ if there exists t_0 such that $F_w(t_0, t_1) \vee S_w(t_0, t_1) = e$.
- succeeds is defined as inevitable $\bigcirc$ ($succeeded(e)$).
- fails is defined as inevitable $\bigcirc$ ($failed(e)$).
- does is defined as inevitable $\bigcirc$ ($done(e)$).

This discussion focuses on single-agent, non-parallel actions. To accommodate parallel actions among multiple agents, the functions must be extended to map to sets of event-agent pairs, indicating which events are performed by which agents.

3.2.4 Semantics of Beliefs, Desires, and Intentions

The traditional view of belief systems sees each world as a set of propositions, with belief represented by a relation between these worlds. A formula is considered believed if it holds true in all worlds accessible from a given world.

In contrast, Cohen and Levesque conceptualize possible worlds as branching timelines of events, offering agents multiple possible paths. Belief in a proposition at a given time means it is true in all worlds accessible from that time. The belief relation is dynamic, allowing for changes in an agent's beliefs over time.

The modal operator **GOAL** reflects an agent's desired situations, similar to belief. Intentions are viewed as chosen future paths, mapped by an accessibility relation. Goal-accessible worlds represent desired states within belief-accessible worlds, while intention-accessible worlds represent chosen paths within goal-accessible worlds.

Transitioning between belief, goal, and intention-accessible worlds reflects shifts in an agent's desires and commitments over time.

- $M, v, w_t \models \mathbf{BEL}(\phi)$ if $\forall w' \in B_t^w$, t_0 such that $M, v, w'_t \models \phi$.
- $M, v, w_t \models \mathbf{GOAL}(\phi)$ if $\forall w' \in G_t^w$, t_0 such that $M, v, w'_t \models \phi$.
- $M, v, w_t \models \mathbf{INTEND}(\phi)$ if $\forall w' \in I_t^w$, t_0 such that $M, v, w'_t \models \phi$.

3.2.5 Axioms for BDI architecture

Belief-Goal If the agent adopts an O-formula α as a goal, the agent believes that the formula

$$\mathbf{GOAL}(\alpha) \supset \mathbf{BEL}(\alpha)$$

This axiom ensures that if the agent aims for a certain outcome, they also hold the belief that it is achievable. However, it is important to note that this doesn't guarantee the agent's belief in reaching that outcome due to the branching nature of time.

$$\forall w' \in B_t^w \ \exists w'' \in G_t^w \quad \text{such} \quad \text{that} \quad w'' \subseteq w'.$$

Strong realism enforces that for every belief-accessible world, there exists a corresponding goal-accessible world that is a sub-world of the belief world. This condition, CI, ensures a strong correspondence between beliefs and goals, indicating a high degree of realism in the agent's reasoning.

Goal-Intention If the agent adopts an O-formula α as intention, the agent should have adopted that formula as a goal to be achieved.

$$\mathbf{INTEND}(\alpha) \supset \mathbf{GOAL}(\alpha)$$

if the agent intends a, she also believes in a. For instance, if the agent intends to perform an event e, she both sets a goal to optionally perform e and believes that she will optionally do e. Nested intentions have interesting implications. For example, if the formula $\mathbf{INTEND}(\textit{inevitable}(\Diamond\mathbf{INTEND}(\text{does}(e))))$ is true, then $\mathbf{BEL}(\textit{optional}(\Diamond\mathbf{INTEND}(\text{does}(e))))$ is true, and so is $\mathbf{BEL}(\text{optional}(\Diamond\mathbf{BEL}(\text{does}(e))))$.

$$\forall w' \in G_t^w \; \exists w'' \in I_t^w \;\; \text{such} \quad \text{that} \;\; w'' \subseteq w'.$$

Similar to the semantic condition of Belief-Goal, we have a semantic condition, which requires that for every goal-accessible world, there exists a corresponding intention-accessible world where the intention-world is a sub-world of the goal-world. This condition ensures a close relationship between goals and intentions.

Intentions-Actions The Axiom captures volitional commitment[Bratman, 1987] by stating that an agent will act if she has an intention towards a single primitive action c. This implies that the agent is committed to attempting the action, but it does not guarantee the success of the action; success depends on the environment.

$$\mathbf{INTEND}(\text{does}(e)) \supset \text{does}(e)$$

whenever an agent intends to perform a specific primitive action, she will indeed carry out that action. However, the axiom does not restrict the agent from performing actions that were not intended, nor does it address non-primitive actions or nested intentions.

Beliefs-Intentions If an agent has a goal to achieve ϕ, the agent believes that she has such a goal.

$$\mathbf{INTEND}(\phi) \supset \mathbf{BEL}(\mathbf{INTEND}(\phi))$$
$$\forall w' \in B_t^w \text{ and } \forall w'' \in G_t^w \;\; \text{we} \quad \text{have} \;\; w'' \subseteq B_t^{w\prime}.$$

Beliefs-Goals If the agent has a goal to achieve ϕ, the agent believes that she has such a goal.

$$\mathbf{GOAL}(\phi) \supset \mathbf{BEL}(\mathbf{GOAL}(\phi))$$
$$\forall w' \in B_t^w \text{ and } \forall w'' \in G_t^w \;\; \text{we} \quad \text{have} \;\; w'' \subseteq B_t^{w\prime}.$$

Goals-Intentions If an agent intends to achieve ϕ, the agent must have the goal to intend ϕ.

$$\mathbf{INTEND}(\phi) \supset \mathbf{GOAL}(\mathbf{INTEND}(\phi))$$
$$\forall w' \in G_t^w \text{ and } \forall w'' \in I_t^w \;\; \text{we have} \;\; w'' \subseteq G_t^{w\prime}.$$

Awarness of primitive events The agent should be aware of all primitive events occurring in the world.

$$\text{done}(e) \supset \mathbf{BEL}(\text{done}(e))$$

No Infinite Deferral if an agent forms an intention, she will eventually give it up.

$$\mathbf{INTEND}(e) \supset \mathit{inevitable}\Diamond(\neg\mathbf{INTEND}(\phi)$$

3.2.6 Proprieties of The Logic

These proprieties are intended to show the relationship between an agent's beliefs, goals, and intentions. The agent believes all valid formulas, intends all valid formulas, and has them as a goal.

In the literature, two important aspects of how beliefs, goals, and intentions interact have been discussed. First, if an agent believes something is always true in all possible future scenarios, they shouldn't necessarily make it a goal or intention. Second, if an agent believes something will always be true and they intend to do something, they shouldn't be forced to intend all the side effects. Based on this, we have the following propositions.

Proposition 1 *A modal formula does not imply a stronger modal formula. If an agent believes something (or inevitably always believes it), it doesn't necessarily mean that the agent also adopts it as a goal.*

- $\boldsymbol{BEL}(\phi) \wedge \neg\boldsymbol{GOAL}(\phi)$
- $\mathit{inevitable}(\Box\boldsymbol{BEL}(\phi)) \wedge \neg\boldsymbol{GOAL}(\phi)$

Proof : For formulas of the form $\mathit{optional}(\psi)$, Case (a) is straightforwardly satisfied. This is because every goal-accessible world is a subset of its corresponding belief-accessible world, and therefore there may not exist any path in which ψ is true.

For other cases, such as when ψ is a first-order formula or a formula of the form $\mathit{inevitable}(\psi)$, we exploit the possibility that there can be goal-accessible worlds that are not belief-accessible. If in one such world, ψ is not true, then the agent will not adopt ψ as a goal. This demonstrates the satisfiability of Case (a). The satisfiability of Case (b) follows a similar reasoning pattern.□

For example, if someone believes the earth is always round, it shouldn't mean they have to make it a goal or intention. Similarly, even if they believe something will eventually happen, like the sun rising, it shouldn't automatically become a goal or intention.

Proposition 2 *A modal operator is not closed under implication with respect to a weaker modality.*

- $\boldsymbol{GOAL}(\phi) \wedge \boldsymbol{BEL}(\textit{inevitable}(\Box(\phi \supset \gamma))) \wedge \neg \boldsymbol{GOAL}(\gamma)$
- $\boldsymbol{GOAL}(\phi) \wedge \textit{inevitable}(\Box \boldsymbol{BEL}(\textit{inevitable}(\Box(\phi \supset \gamma))) \wedge \neg \boldsymbol{GOAL}(\gamma)$

Proof : For every belief-accessible world, there must be a corresponding goal-accessible world. However, the goal relation can extend to worlds that do not necessarily correspond to any belief-accessible world. Therefore, if in one such world, the formula ϕ is not true, the agent will not consider ϕ as a goal. This demonstrates the satisfiability of Case (a). The satisfiability of Case (b) follows a similar pattern.

Both propositions above address the stronger form of beliefs, namely, the inevitability of the agent always believing in ϕ. They can be adjusted accordingly to accommodate the weaker form as well.

It is important to note that despite the propositions mentioned, the agent's goals and intentions adhere to implication closure. In other words, the following formulas are valid within our system:

- $\mathbf{INTEND}(\phi) \wedge \mathbf{INTEND}(\phi \supset \gamma) \supset \mathbf{INTEND}(\gamma)$.
- $\mathbf{GOAL}(\phi) \wedge \mathbf{GOAL}(\phi \supset \gamma) \supset \mathbf{GOAL}(\gamma)$.

□

For instance, if someone plans to go to the dentist but believes it will always cause pain, they shouldn't have to intend the pain. The formalism meets these requirements by allowing agents to have beliefs without them necessarily becoming goals or intentions. We define a relation among modal operators such that belief is weaker than goal, which is weaker than intention. This allows us to compare the strength of different modal formulas.

Chapter 4

Algebraic Logic of Desire

Algebraic logic presents a compelling alternative to traditional modal and syntactic approaches in desire representation due to its unique strengths. It achieves a balance between the expressive power of first-order theories and the modal logic. Incorporating propositions as first-class entities. This allows for richer representation while avoiding self-referential paradoxes that appear in syntactic approaches. In this chapter, we delve into the formalisms and semantics of algebraic logic within the Log_AC_n language [5]. We explore the foundational concepts, syntax, and semantics that underpin the representation of Desire and value judgments within this logical system.

4.1 The Language of Log_AC_n

4.1.1 Formal Syntax

Log_AC_n is algebraic in that it is a language of only terms, some of which denote propositions. The syntax of Log_AC_n, a many-sorted alphabet comprises syncategorematic punctuation symbols and denoting symbols from syntactic sorts σ which is the smallest set containing all of the following sorts [5]

- σ_P.
  the set of terms denoting propositions.
- σ_I.
  the set of terms denoting anything else.
- $\tau_1 \to \tau_2$, for $\tau_1 \in \{\sigma_P, \sigma_I\}$ and $\tau_2 \in \sigma$.
  the syntactic sort of function symbols that takes a single argument of sort σ_p or σ_I and produce a functional term of sort τ_2.

A Log_AC_n alphabet is a union of four disjoint sets: $\Omega \cup \Xi \cup \Sigma \cup \Lambda$.

- Ω: *Signature of the language, It is a non-empty, countable set of constant and function symbols.*

- Ξ : Set $\Xi = \{x_i, a_i, p_i\}_{i\in\mathbb{N}}$ is a countably infinite set of variables, where $x_i \in \sigma_I$, $a_i \in \sigma_A$, and $p_i \in \sigma_P$, for $i \in \mathbb{N}$.

- Σ: Set of syncategorematic symbols, including the comma, various matching pairs of brackets and parentheses, and the symbol $\forall$.

- Λ: Set of logical symbols including $(=)$, $(\neg)$, $(\wedge)$, $(\vee)$, belief (**B**), intention (**I**), goodness (**Good**), badness (**Bad**), comparison of goodness (**Btr**), and comparison of position on a scale (**Prj**).

  $\bigcup_{i=1}^{n}\{Good_i, Bad_i\} \subseteq \sigma_A \to \sigma_P \to \sigma_P$.

  $\bigcup_{i=1}^{n}\{Btr_i\} \subseteq \sigma_A \to \sigma_P \to \sigma_P \to \sigma_P$.

L_Ω represents the set of terms in the Log_AC_n language with a given signature Ω. These terms are formed according to specific rules outlined below:

1. The set of Variables $\Xi \subset L_\Omega$.

2. $c \in L_\Omega$, where $c \in \Omega$ is a constant symbol.

3. $f(t_1, \ldots, t_n) \in L_\Omega$, where $f \in \Omega$ is of type $\tau_1 \to \cdots \to \tau_n \to \tau$ $(n > 0)$ and t_i is of type τ_i.

4. $\neg t \in L_\Omega$, where $t \in \sigma_P$.

5. $(t_1 \otimes t_2) \in L_\diamond$, where $\otimes \in \{\wedge, \vee\}$ and $t_1, t_2 \in \sigma_P$.

6. $(t_1 \otimes t_2) \in L_\Omega$, where $\otimes \in \{\wedge, \vee\}$ and $t_1, t_2 \in \sigma_P$.

7. $\{B(t_1, t_2), I(t_1, t_2), \mathrm{Good}_i(t_1, t_2), \mathrm{Bad}_i(t_1, t_2)\} \in L_\diamond$, where $t_1 \in \sigma_A$, $t_2 \in \sigma_P$, and $1 \leq i \leq n$.

8. $Btr_i(t_1, t_2, t_3) \in L_\Omega$, where $t_1 \in \sigma_A$, $t_2 \in \sigma_P$, $t_3 \in \sigma_P$, and $1 \leq i \leq n$.

9. $Prj_i^j(t_1, t_2) \in L_\Omega$, where $t_1 \in \sigma_A$ and $t_2 \in \sigma_P$, for $1 \leq i < j \leq n$.

As usual, terms involving $\Rightarrow$, $\Leftrightarrow$, and $\exists$ may be introduced as abbreviations in the standard way.

4.1.2 Semantics

In This section, we shall provide some of the logical semantics which will be helpful in studying desire properties

Definition 1 *A Log_AC_n structure with the tuples $(\mathfrak{C} = \mathfrak{D}, \mathfrak{U}, \mathfrak{S})$ where*

- $\mathfrak{D}$ *the domain of discourse, is a set with two disjoint, non-empty, countable subsets $\mathcal{P}$ and $\mathcal{A}$.*
- $\mathfrak{U} = \mathfrak{p}, +, \ . \ , -, \top, \bot$
- $\mathfrak{S}$ *is a set of n scales, where every scale is a triple $\langle \succ_i, \ddot{\smile}_i, \ddot{\frown}_i \rangle$*
  - $\succ_i : \mathcal{A} \times \mathcal{P} \times \mathcal{P} \to \mathcal{P}$.
  - $\ddot{\smile}_i : \mathcal{A} \times \mathcal{P} \to \mathcal{P}$.
  - $\ddot{\frown}_i : \mathcal{A} \times \mathcal{P} \to \mathcal{P}$.

  Moreover, each scale is constrained as follows, for every $p_1, p_2, p_3 \in P$ and $a \in A$. (For better readability, we will henceforth write "$(p_1 \succ_i^a p_2)$" for "$\succ_i (a, p_1, p_2)$".)
  $\mathfrak{S}1$. $(p_1 \succ_i^a p_2) \cdot (p_2 \succ_i^a p_1) = \bot$.
  $\mathfrak{S}2$. $(p_1 \succ_i^a p_2) \cdot -(p_3 \succ_i^a p_2) \cdot -(p_1 \succ_i^a p_3) = \bot$.
  $\mathfrak{S}3$. $\ddot{\smile}_i (a, p_1) \cdot - \ddot{\smile} i(a, p_2) \cdot -(p_1 \succ_i^a p_2) = \bot$.
  $\mathfrak{S}4$. $\ddot{\frown}_i (a, p_1) \cdot - \ddot{\frown} i(a, p_2) \cdot -(p_2 \succ_i^a p_1) = \bot$.
  $\mathfrak{S}5$. $\ddot{\smile}_i (a, p_1) \cdot - \ddot{\frown} i(a, p_1) = \bot$.

In our model, the domain $\mathfrak{D}$ is divided into two main components: a set $\mathcal{P}$ of propositions forming a Boolean algebra, and a set $\overline{\mathcal{P}}$ of individuals which includes a non-empty set A representing agents. Additionally, we have a set $\mathfrak{S}$ consisting of n scales, each representing a preference structure in the language we are using.

$\mathfrak{S}1$ and $\mathfrak{S}2$ show that an agent's ordering of propositions on a scale is asymmetric, transitive, and total. $\mathfrak{S}3$ and $\mathfrak{S}4$ show that good propositions are more preferred than propositions that are not good and that bad propositions are less preferred than propositions that are not bad. This maintains the opposition relation between good and bad on a scale. The constraint $\mathfrak{S}5$ ensures that the two poles, representing the judgments of propositions as good or bad, do not coincide. In other words, it asserts that a proposition cannot be simultaneously evaluated as both good and bad by the same agent on the same scale. The set $\mathfrak{P}$ of projections establishes order-preserving isomorphism between each pair of scales. In simpler terms, projections allow us to compare how agents evaluate propositions across different contexts or situations. order-preserving isomorphism describes a mapping between preference structures that preserves the ordering of propositions. If proposition A is preferred over proposition B in one scale, the same preference relationship holds true after applying the projection to another scale.

Definition 2 *Let $\alpha \in \sigma_A$, $\phi, \psi \in \sigma_P$, and $S \cup \{i, j\} \subseteq \{1, \ldots, n\}$ with $i < j$.*

1. $\boldsymbol{Des}_i(\alpha, \phi) =_{def} \boldsymbol{Good}_i(\alpha, \phi) \wedge \boldsymbol{Btr}_i(\alpha, \phi, \neg\phi)$
2. $\boldsymbol{Des}_S(\alpha, \phi) =_{def} \bigvee_{k \in S} \boldsymbol{Des}_k(\alpha, \phi)$
3. $\boldsymbol{Des}(\alpha, \phi) =_{def} \bigvee_{i=1}^{n} \boldsymbol{Des}_i(\alpha, \phi)$
4. $\boldsymbol{DES}_S(\alpha, \phi) =_{def} \bigwedge_{k \in S} \boldsymbol{Des}_k(\alpha, \phi)$
5. $\boldsymbol{DES}(\alpha, \phi) =_{def} \bigwedge_{i=1}^{n} \boldsymbol{Des}_i(\alpha, \phi)$
6. $\boldsymbol{GdAndDn}_i(\alpha, \phi, \psi) =_{def} \boldsymbol{Good}_i(\alpha, \phi) \wedge \boldsymbol{Good}_i(\alpha, \psi) \Rightarrow [\boldsymbol{Good}_i(\alpha, \phi \wedge \psi)]$
7. $\boldsymbol{GdOrDn}_i(\alpha, \phi, \psi) =_{def} \boldsymbol{Good}_i(\alpha, \phi) \vee \boldsymbol{Good}_i(\alpha, \psi) \Rightarrow [\boldsymbol{Good}_i(a, p \vee \psi)]$
8. $\boldsymbol{LG}_i(\alpha, \phi) =_{def} \forall p[(\phi \wedge p) = \phi \Rightarrow (\boldsymbol{Good}_i(\alpha, \phi) \Rightarrow \boldsymbol{Good}_i(\alpha, p))]$
9. $\boldsymbol{LG}(\alpha, \phi) =_{def} \bigwedge_{i=1}^{n} \boldsymbol{LG}_i(\alpha, \phi)$

Proposition 1 *For any $Log_A C_n$ language, the following is true, for $1 \leq i < j \leq n$:*

1. $\models \boldsymbol{Btr}_i(a, p_1, p_2) \Rightarrow \neg\boldsymbol{Btr}_i(a, p_2, p_1)$
2. $\models \boldsymbol{Btr}_i(a, p_1, p_2) \wedge \boldsymbol{Btr}(a, p_2, p_3) \Rightarrow \boldsymbol{Btr}_i(a, p_1, p_3)$
3. $\models \neg\boldsymbol{Btr}_i(a, p, p)$
4. $\models \boldsymbol{Btr}_i(a, p_1, p_2) \Rightarrow \neg\boldsymbol{Btr}_j(a, Pr_{ij}(a, p_2), Pr_{ij}(a, p_1))$
5. $\models \boldsymbol{Good}_i(a, p_1) \wedge \neg\boldsymbol{Good}_i(a, p_2) \Rightarrow \boldsymbol{Btr}_i(a, p_1, p_2)$

4.1.3 Algebraic Logic of Desire

The choice of that notion of desire comes in handy to avoid non-monotonicity problems regarding desire because classically desire is not always monotonic. Adding new information or beliefs can change what we desire, even if the original object of desire remains unchanged. For example, learning that a desired food is unhealthy might diminish the desire for it. however, the introduction of scales makes it possible to desire and undesire the same thing within different scales. (see more [4]) In this section, we shall delve into some of the properties of desire according to the notion of desire represented in $Log_A C_n$ [5] and will try to prove some of them. But before we begin we shall define some definitions that will help us prove certain properties.

Definition 3 *Let $\alpha \in \sigma_A$, $\phi, \psi \in \sigma_P$, and $S \cup \{i, j\} \subseteq \{1, \ldots, n\}$ with $i < j$.*

1. $\boldsymbol{BtrAndDn}_i(\alpha, \phi, \psi) =_{def} \boldsymbol{Btr}_i(\alpha, \phi, \neg\phi) \wedge \boldsymbol{Btr}_i(\alpha, \psi, \neg\psi) \Rightarrow [\boldsymbol{Btr}_i(\alpha, \phi \wedge \psi, \neg(\phi \wedge \psi))]$
2. $\boldsymbol{BtrOrDn}_i(\alpha, \phi, \psi) =_{def} \boldsymbol{Btr}_i(\alpha, \phi, \neg\phi) \vee \boldsymbol{Btr}_i(\alpha, \psi, \neg\psi) \Rightarrow [\boldsymbol{Btr}_i(\alpha, (\phi \vee \psi), \neg(\phi \vee \psi))]$
3. $\boldsymbol{LBtr}_i(\alpha, \phi) =_{def} \forall\psi[(\phi \wedge \psi) = \phi \Rightarrow (\boldsymbol{Btr}_i(\alpha, \phi) \Rightarrow \boldsymbol{Btr}_i(\alpha, \psi))]$

We define this condition for the purpose of having one possible algebraic logic of desire, they are context-dependent and restricted. So generalizations of the above definitions are generally not advisable.

This definition allows us to prove some propositions like closure under implication for desire, which creates paradoxes in different frameworks like modal logic. (see Prior's good Samaritan paradox [9])

Proposition 2 *Let* $S \cup \{i, j, k\} \subseteq \{1, \ldots, n\}$, *with* $|S| > 1$ *and* $i < j < k$.

$$\models \boldsymbol{Des}_i(a, p) \Rightarrow \neg\boldsymbol{Des}_i(a, \neg p)$$

Proof.

1. $\mathbf{Des}_i(a, p) \wedge \mathbf{Des}_i(a, \neg p)$ (Assumption).
2. $\mathbf{Good}_i(a, p) \wedge \mathbf{Btr}_i(a, p, \neg p)$ (1, Definition 2).
3. $\mathbf{Good}_i(a, \neg p) \wedge \mathbf{Btr}_i(a, \neg p, p)$ (1, Definition 2).
4. $\mathbf{Btr}_i(a, p, \neg p)$ (3, $\wedge$-Elimination).
5. $\mathbf{Btr}_i(a, \neg p, p)$ (4, $\wedge$-Elimination).
6. $\mathbf{Btr}_i(a, p, \neg p) \Rightarrow \neg\mathbf{Btr}_i(a, \neg p, p)$ (Proposition 1)
7. $\neg(\mathbf{Des}_i(a, p) \wedge \mathbf{Des}_i(a, \neg p))$
8. $\mathbf{Des}_i(a, p) \Rightarrow \neg\mathbf{Des}_i(a, \neg p)$ (7, De Morgan's Law) $\square$

According to the propositions of $\mathbf{Btr}_i$ in Log_AC_n specifically 1.1(1) you cannot find proposition $\mathbf{p}$ better than $\neg\,\mathbf{p}$ and $\neg\,\mathbf{p}$ better than $\mathbf{p}$ on the same scale.

Since our initial assumption led to a contradiction, the assumption must be false, therefore we concluded the negation of our initial assumption.

This proposition states that if agent a desires proposition p on a certain scale i, then it does not desire the negation of p on the same scale. It also shows us on scale (i) the comparative nature of desire. We desire things more or less than others. Consider an agent, Alice (a), Alice evaluates her desire to attend music festivals (p) on a specific scale (i), representing the lineup quality of the festivals.

- Alice desires to attend music festivals (p) considering the lineup quality represented by scale (i):

$$\mathbf{Des}_i(a, p) \quad \text{where } i \text{ is Alice's scale}$$

- However, Alice does not desire not to attend music festivals ($\neg p$) on the same scale (i):

$$\neg\mathbf{Des}_i(a, \neg p) \quad \text{where } i \text{ is Alice's scale}$$

In this scenario, Alice desires to attend music festivals (p) on her scale (i), and she does not desire not to attend music festivals ($\neg\, p$) on the same scale (i) because she considers the lineup quality to be significant in her decision-making process. Thus, Alice's preferences align with **proposition 1.2**(1), where she desires (p) on her scale i but does not desire ($\neg p$) on the same scale (i). It could be the case that Alice desires to not attend music festivals on a different scale, for example, the scale of cost.

Proposition 3 *Let $S \cup \{i, j, k\} \subseteq \{1, \ldots, n\}$, with $|S| > 1$ and $i < j < k$.*

$$\models (\boldsymbol{GdOrDn}_i(a, p, q) \wedge \boldsymbol{BtrOrDn}_i(a, p, q)) \wedge (\boldsymbol{Des}_i(a, p) \vee \boldsymbol{Des}_i(a, q)) \Rightarrow \boldsymbol{Des}_i(a, p \vee q)$$

Proof.

1. $\mathbf{Des}_i(a, p) \vee \mathbf{Des}_i(a, q)$ (Assumption)
2. $\mathbf{Good}_i(a, p) \wedge \mathbf{Btr}_i(a, p, \neg p)$ (1, Definition 2)
3. $\mathbf{Good}_i(a, q) \wedge \mathbf{Btr}_i(a, q, \neg q)$ (1, Definition 2)
    4. $\mathbf{BtrOrDn}_i(a, p, q)$ (Assumption)
        5. $\mathbf{GdOrDn}_i(a, p, q)$ (Assumption)
        6. $\mathbf{Good}_i(a, p)$ (4, $\wedge$-Elimination)
        7. $\mathbf{Good}_i(a, q)$ (5, $\wedge$-Elimination)
        8. $\mathbf{Good}_i(a, p \vee q)$ (8, 10, Definition 2)
    9. $\mathbf{Btr}_i(a, p, \neg p)$ (2, $\wedge$-Elimination)
    10. $\mathbf{Btr}_i(a, q, \neg q)$ (3, $\wedge$-Elimination)
    11. $\mathbf{Btr}_i(a, p \vee q, \neg(p \vee q))$ (9, 10, Definition 3)
12. $\mathbf{Good}_i(a, p \vee q) \wedge \mathbf{Btr}_i(a, p \vee q, \neg(p \vee q))$(8, 11, $\wedge$-Introduction)

13. $\mathbf{Des}_i(a, p \vee q)$ (12, Definition 2) $\square$

The proposition states that if an agent either desires proposition (p) or desires proposition q on the scale (i), and if (p) is evaluated as good and better than ($\neg p$), while (q) is also evaluated as good and better than ($\neg q$) on the scale (i), then the agent must desire the disjunction (p $\vee$ q) on the same scale (i).

Suppose Alice either desires to eat ice cream (p) or desires to listen to her favorite music (q) on some evaluation scale i. Additionally, Alice evaluates eating ice cream as a good experience that is preferable to not eating ice cream ($\neg$p), and similarly, she evaluates listening to her favorite music as good and preferable to not listening to it ($\neg$ q), both on a scale i. Then, according to the proposition, Alice must also desire the disjunctive experience of either eating ice cream or listening to her favorite music (p $\vee$ q) on the scale i. The reasoning is that since Alice finds eating ice cream to be good and better than not eating it, and also finds listening to music to be good and better than not listening, then by the definitions of the **GdOrDn** and **BtrOrDn** operators, the disjunction (p $\vee$ q) is evaluated as a good experience that is preferable to its negation $\neg$(p $\vee$ q) on scale (i). Given Alice desires at least one of the experiences (p or q) and evaluates their disjunction as preferable, the desire definition implies Alice must intrinsically desire the disjunctive experience (p $\vee$ q) of either eating ice cream or listening to music on that same scale i. So if Alice desires ice cream or desires music, and broadly evaluates both as positive experiences, then she should naturally also desire the possibility of having one or the other occur or maybe both of them.

Proposition 4 *Let $S \cup \{i, j, k\} \subseteq \{1, \ldots, n\}$, with $|S| > 1$ and $i < j < k$.*

$$\models (\boldsymbol{GdAndDn}_i(a, p, q) \wedge \boldsymbol{BtrAndDn}_i(a, p, q)) \wedge (\boldsymbol{Des}_i(a, p) \wedge \boldsymbol{Des}_i(a, q)) \Rightarrow \boldsymbol{Des}_i(a, p \wedge q)$$

Proof.

1. $\mathbf{Des}_i(a, p) \wedge \mathbf{Des}_i(a, q)$ (Assumption)
2. $\mathbf{Good}_i(a, p) \wedge \mathbf{Btr}_i(a, p, \neg p)$ (1, Definition 2)
3. $\mathbf{Good}_i(a, q) \wedge \mathbf{Btr}_i(a, q, \neg q)$ (1, Definition 2)
    4. $\mathbf{BtrAndDn}_i(a, p, q)$ (Assumption)
        5. $\mathbf{GdAndDn}_i(a, p, q)$ (Assumption)
        6. $\mathbf{Good}_i(a, p)$ (2 and $\wedge$-Elimination)
        7. $\mathbf{Good}_i(a, q)$ (3 and $\wedge$-Elimination)
        8. $\mathbf{Good}_i(a, p \wedge q)$ (6, 7, Definition 2)

9. $\mathbf{Btr}_i(a, p, \neg p)$ (2 and $\wedge$-Elimination)

10. $\mathbf{Btr}_i(a, q, \neg q)$ (3 and $\wedge$-Elimination)

11. $\mathbf{Btr}_i(a, p \wedge q, \neg(p \wedge q))$ (9, 10 Definition 3)

12. $\mathbf{Good}_i(a, p \wedge q) \wedge \mathbf{Btr}_i(a, p \wedge q, \neg(p \wedge q))$ (8, 11 and $\wedge$-Introduction)

13. $\mathbf{Des}_i(a, p \wedge q)$ (12, Definition 2) $\square$

This proposition means that if an agent desires proposition(p) on scale i and he also desires proposition (q) on the same scale i.

Consider Alice, who evaluates experiences based on how pleasurable they are. Alice enjoys eating ice cream(p) and finds it preferable to not eating ice cream($\neg p$). She also enjoys listening to her favorite music (q) and prefers it to not listening to music($\neg q$). then by the definitions of the **GdAndDn** and **BtrAndDn** operators, the conjunction (p $\wedge$ q) is a good experience that is preferable to its negation $\neg$(p $\wedge$ q) on the scale (i). It follows that Alice would likely enjoy the combined experience of eating ice cream while listening to her favorite music even more (and would prefer it to not doing either).

Alice would never say that she desires the conjunction of (p) and (q) unless she finds both good and both propositions are better than their negation.

Proposition 5 *Let $S \cup \{i, j, k\} \subseteq \{1, \ldots, n\}$, with $|S| > 1$ and $i < j < k$.*

$$\models \boldsymbol{DES}_S(a, p) \Rightarrow \neg \boldsymbol{DES}_S(a, \neg p)$$

Proof.

1. $\mathbf{DES}_S(a, p) \wedge \mathbf{DES}_S(a, \neg p)$(Assumption).

2. $(\bigwedge_{k \in S} \mathbf{Des}_k(a, p)) \wedge (\bigwedge_{k \in S} \mathbf{Des}_k(a, \neg p))$.(1, Definition 2)

3. $\mathbf{Des}_k(a, p) \wedge \mathbf{Des}_k(a, \neg p)$ (2, $\wedge$-Elimination)

4. $\mathbf{Good}_k(a, p) \wedge \mathbf{Btr}_k(a, p, \neg p)$(3, Definition 2).

5. $\mathbf{Good}_k(a, \neg p) \wedge \mathbf{Btr}_k(a, \neg p, p)$(3, Definition 2).

6. $\mathbf{Btr}_k(a, p, \neg p)$(4, $\wedge$-Elimination)

7. $\mathbf{Btr}_k(a, \neg p, p)$(5, $\wedge$-Elimination)

8. $\mathbf{Btr}_i(a, p, \neg p) \Rightarrow \neg \mathbf{Btr}_i(a, \neg p, p)$(Proposition 1)

9. $\neg(\mathbf{Des}_k(a,p) \wedge \mathbf{Des}_k(a,\neg p))(1,\ 8)$

10. $\mathbf{Des}_k(a,p) \Rightarrow \neg\mathbf{Des}_k(a,\neg p)(9,$ De Morgan's Law)

11. $\bigwedge_{k\in S} \mathbf{Des}_k(a,p) \Rightarrow \neg \bigwedge_{k\in S} \mathbf{Des}_k(a,\neg p)$ (10, $\wedge$-Introduction)$\square$

This proposition states that if agent a desires proposition p on all scales within a given set of scales, then it does not desire the negation of p on all of the scales in the same set. We fall into the same contradiction we reached in Proposition 1.2(1) Because k was arbitrary, this contradiction holds for any element within the set S. If it fails for one desire in the set, the entire $\mathbf{DES}_S$ definition becomes false.

This is followed by the intuition that Alice desires to attend music festivals(p) across different scales, representing different aspects.The set S contains scales cost,lineup quality, location

- Alice desires to attend music festivals (p) on all scales within the set (S).This means that Alice is considering factors such as cost, lineup quality, and location.:

$$\mathbf{DES}_S(a,p) \quad \text{where } S \text{ is Alice's set of scales}$$

- However, Alice does not desire not to attend music festivals ($\neg p$) on all of the scales within the set (S):

$$\mathbf{DES}_S(a,\neg p) \quad \text{where } S \text{ is Alice's set of scales}$$

  She desires to attend music festivals (p) on all scales within the set S, and she does not desire not to attend music festivals ($\neg p$) on all of the scales within the set S.

Proposition 6 *Let $S \cup \{i,j,k\} \subseteq \{1,\ldots,n\}$, with $|S| > 1$ and $i \neq j \neq k$.*

$$\not\models \boldsymbol{Des}_S(a,p) \Rightarrow \neg\boldsymbol{Des}_S(a,\neg p)$$

This proposition means that it is not the case if an agent (a) desires (p) on some scale in the Set (S), the agent does not desire ($\neg p$) on some scale in the Set (S), the notion of $\mathbf{Des}_S$ is different of $\mathbf{DES}_S$ according to definition 2.

Proof.

We shall prove this proposition using a counterexample. Assume an agent Alice (a) desires "attending music festivals" (p) and we have the Set (S) containing Cost, Location, Alice is on a tight budget, and expensive music festivals are not desirable to her. This means $\mathbf{Des}_{Cost}(a,\neg p)$ is true. (1) However, Alice loves music festivals and would greatly enjoy attending one if it is held in a convenient location. This implies $\mathbf{Des}_{Location}(a,p)$ is also true. (2) The implication is false because we have the assumption ($\mathbf{Des}_{Cost}(a,p)$) holding true, but the conclusion of the formula ($\neg\mathbf{Des}_{Location}(a,p)$) is false.

Proposition 7 *Let $S \cup \{i, j, k\} \subseteq \{1, \ldots, n\}$, with $|S| > 1$ and $i < j < k$.*

$$\models (\boldsymbol{LG}_i(a, p) \wedge \boldsymbol{LBtr}_i(a, p, q)) \wedge (\boldsymbol{Des}_i(a, p) \wedge (\forall p, q(p \wedge q) = p) \Rightarrow \boldsymbol{Des}_i(a, q)$$

Proof.

1. $\forall p, q(p \wedge q) = p$ (Hypothesis)
   2. $\mathbf{Des}_i(a, p)$(Assumption)
   3. $\mathbf{Good}_i(a, p) \wedge \mathbf{Btr}_i(a, p, \neg p)$ (2, Definition 2)
      4. $\mathbf{LBtr}_i(a, p, q)$ (Assumption)
         5. $\mathbf{LG}_i(a, p)$ (Assumption)
         6. $\mathbf{Good}_i(a, p)$ (3, $\wedge$-Elimination)
         7. $\mathbf{Good}_i(a, q)$ (1, 6, Definition 2)
      8. $\mathbf{Btr}_i(a, p, \neg p)$ (3, $\wedge$-Elimination)
      9. $\mathbf{Btr}_i(a, q, \neg q)$ (1, 8, Definition 3)
   10. $\mathbf{Good}_i(a, q) \wedge \mathbf{Btr}_i(a, q, \neg q)$ (7, 9, $\wedge$-Introduction)
11. $\mathbf{Des}_i(a, q)$ (10, Definition 2) $\square$

This proposition states that if an agent considers all logical consequences of a proposition to be good ($\mathbf{LG}_i$) and better than their negations ($\mathbf{LBtr}_i$), then desiring the original proposition (p) also leads to desiring any proposition (q) that is logically equivalent to p.

The proposition establishes a form of closure under logical equivalence for desire within the $Log_A C_n$ language. It demonstrates that an agent's desires are not only influenced by their individual preferences but also by the logical relationships between propositions and their consequences. The closure under logical implication showcases the power of algebraic languages, for example, in modal logic that proposition is historically paradoxical, p implies $(p \vee q)$ then if I desire eating icecream I might desire killing somebody(see more about Ross's paradox in deontic logic [9]).

Chapter 5

Conclusion

This paper contains two main contributions. First, we reviewed together all types of logic of desire, we talked briefly about each one mentioning its history, definition, and problems. We extended the logic of desire in the language $Log_A C_n$, we added new definitions to support our main goal which was proving some widely known properties about desire in algebraic notions.

The implications of this work extend beyond theoretical exploration. The formalizing of desire in algebraic logic could lead to the development of more human-like AI agents that are not only more intelligent but also more aligned with human values and desires.

Chapter 6

Future Work

A common approach in AI planning systems is to represent desires as goals, assuming a binary distinction between desirable (goal-satisfying) and undesirable outcomes. However, this simplification fails to capture the nature of human preferences. For instance, we might desire to have both p and q, but if only one is achievable, we would prefer p.(see more [3])

To address this limitation, we can establish a notion of relative desire within the $Log_A C_n$ language. This involves using scales to rank preferences, allowing us to compare propositions across different scales through projections. In this framework, one proposition is preferred over another if it ranks higher on the relevant scale.

Specifically, if agent α desires proposition ϕ on scale i, and we wish to express this desire relative to scale j, we utilize the projection $\mathbf{Prj}_i^j$. This function maps the value of ϕ from scale i to scale j, thus preserving the order of preference while adjusting the context of evaluation. Another idea is to use the notion of $\mathbf{Btr}_i$ to define a notion of relative desire between different propositions within the same scale.

Lastly, Second-order attitudes are very famous when it comes to psychology literature and second-order desires are ignored throughout the history of psychology and AI research, yet Frankfurt's work pointed out that second-order desires are very important when it comes to the ability of love, care, and having a free will.

We are unsure if AI agents nowadays understand love, and care, or even have free will in their output. We sometimes force intelligent agents to do stuff based on our given input yet, having incomplete knowledge we can orient agents to do stuff in the wrong direction. Maybe giving them free will help improve such problems.

By trying to represent and reason about second-order desires, maybe we can help develop intelligent agents that can care, and love and make decisions based on that, well at least the bare minimum we want them to know what care and love are. So they can understand us when we talk to them in terms of this topics.

Appendix

Appendix A

Brief Description of Modal Logic

Modal logic is a family of formal systems that deal with concepts of necessity (represented by the symbol '$\Box$') and possibility ('$\diamond$'). It builds upon propositional logic by adding axioms and rules specifically for these modal operators.

The basic system, K, includes the following key principles:

- **Necessitation Rule**: If a statement is a theorem of logic, then it's necessarily true.
- **Distribution Axiom**: If it's necessary that 'if A then B', then 'if necessarily A, then necessarily B'.

Different modal logics are created by adding further axioms to K. Some notable systems include:

- **M** (or T): Introduces the axiom that whatever is necessary is true ($\Box A \rightarrow A$).
- S4: Adds the axiom that if something is necessary, it's necessarily necessary ($\Box A \rightarrow \Box\Box A$)
- **S5**: Includes the axiom that if something is possible, it's necessarily possible ($\diamond A \rightarrow \Box \diamond A$).
- **B**: Introduces the axiom $A \rightarrow \Box \diamond A$, though its interpretation requires careful consideration due to potential ambiguities in natural language. itemize those in latex

The choice of axioms depends on the intended use of 'necessarily' and 'possibly,' as these words have various interpretations in different contexts. The various modal logics and their relationships can be understood more deeply through their possible world semantics.

Bibliography

[1] Michael E. Bratman. A desire of one's own. *The Journal of Philosophy*, 100(5):221–242, 2003.

[2] Donald Davidson. *Essays on Actions and Events: Philosophical Essays Volume 1*. Clarendon Press, Oxford, GB, 2001.

[3] Jon Doyle, Yoav Shoham, and Michael Wellman. A logic of relative desire. 04 1994.

[4] Haythem Ismail. Log(a)b: A first-order, non-paradoxical, algebraic logic of belief. *Logic Journal of the IGPL*, 5, 10 2012.

[5] Haythem Ismail. *The Good, the Bad, and the Rational: Aspects of Character in Logical Agents*, pages 139–164. 01 2020.

[6] David Lewis. Desire as belief. *Mind*, 97(387):323–332, 1988.

[7] David Lewis. Desire as belief ii. *Mind*, 105(418):303–313, 1996.

[8] Yongming li, Yali Li, and Zhanyou Ma. Computation tree logic model checking based on possibility measures. *Fuzzy Sets and Systems*, 262, 01 2014.

[9] Paul McNamara and Frederik Van De Putte. Deontic Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[10] Carolyn R. Morillo. The reward event and motivation. *Journal of Philosophy*, 87(4):169–186, 1990.

[11] Anand Srinivasa Rao and Michael P. Georgeff. Modeling rational agents within a bdi-architecture. In *International Conference on Principles of Knowledge Representation and Reasoning*, 1997.

[12] Thomas M Scanlon. *What we owe to each other*. Harvard University Press, 2000.

[13] Tim Schroeder. Desire. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

[14] Timothy Schroeder. *Three Faces of Desire*. Oxford University Press, New York, US, 2004.

[15] Yi Zhou and Xiaoping Chen. Partial implication semantics for desirable propositions. pages 606–612, 01 2004.