Sense-Making Machines

A dissertation

submitted by

Vasanth Sarathy

B.S., University of Arkansas; S.M., Massachusetts Institute of Technology;

J.D., Boston University School of Law

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science: Cognitive Science

TUFTS UNIVERSITY

August 2020

Dissertation Committee

Matthias Scheutz (Supervisor)
Daniel Dennett (Supervisor)
Anselm Blumer
Jivko Sinapov
Kamal Premaratne

To Laura, the love of my life, who has supported me through my intellectual and trivial pursuits

Acknowledgments

I am lucky. Lucky to be supported by a loving family. Lucky to be guided by caring and committed role-models. Lucky to be part of an institution and research community interested in solving problems and helping people. Lucky to have acquired the proper foundations to explore some of the most exciting questions about the mystery of our human mind. Lucky to have had the opportunity to work and be influenced by so many great people. My father-in-law once told me, "the harder you work the luckier you get." I think he was almost right. It always seems to me that the harder *my community* and those around me work, the luckier I get. I am lucky because of them and for that I am truly thankful.

While I could not have written and defended this dissertation without the help of many people throughout my life, I will focus here on those who were particularly influential to me in the last five years as I was writing it. I must begin by thanking my primary advisor, Matthias Scheutz, for guiding me through this process and helping me grow as a researcher. He saw in me what many others did not, and in some sense, what I did not see in myself. I want to thank him for helping me focus my research questions, challenge my thinking and advocating for me at every chance. I have really enjoyed getting in front of a whiteboard with him and discussing how even the simplest examples of human intelligence (things young children can do) contain layers of richness, elusive computational aspects and questions that we humans have yet to understand. I have so many snapshots of these whiteboards filled with red and blue and black marks memorializing our insights, many of which were elaborated in this dissertation. Most important of all, I want to thank Matthias for taking a chance on me, a lawyer with little background in cognitive science interested in exploring the computational principles of intelligence and creativity.

I want to thank my secondary advisor Dan Dennett, for his wisdom, for his enthusiasm for my work, and for his ability to see deeply insightful connections between various themes in my research. I have enjoyed Dan sharing his stories about the field of cognitive science and challenging my thinking with puzzles and observations about various phenomena associated with thinking and intelligence, which are still open problems.

I also want to thank the other members of my dissertation committee: Kamal Premaratne for his kindness and sheer mathematical intelligence – helping me understand the bounds of what it means to be "uncertain" about something and being patient when the mathematics became more challenging to intuit. I finally, want to thank Anselm Blumer and Jivko Sinapov for pushing me to make my dissertation the best it can be, while offering substantial guidance to help me drive that effort.

This work would not be where it is without the various collaborators and mentors I have had, some as co-authors, some as friends, and others as idea-bouncer-offers and gut-checkers. All of them have motivated me to become engaged with my research and teaching community at Tufts and beyond. I want to especially thank Bertram Malle, Chitta Baral, Ayanna Thomas, J.P. de Ruiter, Liz Race, Kelly McLaughlin, Shaun Patel, Ben Hescott, and Patrícia Alves-Oliveira.

Through these five years, the Human-Robot Interaction (HRI Lab) has been my home away from home. I cannot fully summarize the impact my labmates have had on me – pushing my thinking intellectually, reading my first drafts of papers, commiserating over conference rejections, celebrating conference acceptances, listening to my practice talks and humoring me running into their office with a half-baked idea that I think they should hear immediately. I especially want to thank Evan Krause, Brad Oosterveld, and Ravenna Thielstrom, for their help in getting my robots to do what they are supposed to do. I want to thank Thomas Arnold, who has enriched my thinking in so many ways, and particularly helped me realize the intricacy of the ethical fabric supporting humanity's relationship with technology, particular artificial intelligence. I really appreciate that all I had to was swing my chair around and say "Thomas, do you have a minute? I was wondering..." My conversations over the years with Tom Williams, Willie Wilson and Dan Kasenberg helped me shape my views on cognitive science and helped me connect with the

research community in AI more broadly.

Now, this would not have been possible without the love and support of my family abroad: my (late) father who believed in me and taught me the singular importance of integrity, loyalty, honesty and sheer hard work, my mother who showed me the importance of community and the value of giving back, and my brother who has always been on team Vasanth. My family in the U.S., has been my biggest cheerleader and coach – my father-in-law and mother-in-law have taught me the value of believing in myself and in being true to myself. Finally, I'd like to thank Laura, my wife, and to whom I've dedicated this dissertation. She helped me strike a balance between research and home , she supported me through career transitions, family tragedies and of course all while being my tireless advocate and the mother of our children.

They work hard for me.
Therefore, I am lucky.
Q.E.D.

Vasanth Sarathy

TUFTS UNIVERSITY

August 2020

Sense-Making Machines

Vasanth Sarathy

Although statistical machine learning techniques have led to significant advances in AI systems, they are still far from demonstrating fundamental intelligence capabilities possessed by human toddlers and even some animals. After decades of research and millennia of scientific and philosophical thought, the central goals of AI – to explain and replicate human intelligence and creativity – still remain unmet. In this thesis, I argue for instilling in AI systems, the ability to continually "make sense" of its changing world to guide behavior and understand perceptual information. Different from current mainstream AI approaches, I propose that agents maintain and update mental representations of the world that allow them to reason about symbolic concepts under uncertainty. I show that such representations and inference machinery are needed at all levels of cognitive processing – from language interpretation, basic visual perception and action selection to high-level deliberation and even creative problem-solving. In the latter, sense-making becomes sense-breaking, in which I demonstrate how an agent can break its own assumptions and biases in order to discover novel ideas and solutions. Symbolic representations allow an agent to reason beyond statistical patterns, verify the veracity of their knowledge, recognize gaps in their understanding, raise questions, and explore the world to seek out answers. In doing so, such representations also allow artificial agents to provide us, humans, explanations of their behaviors, allow us to better interpret and understand their actions, and ensure that they comply with our social norms.

Contents

Acknowledgments iii

Abstract vi

List of Tables xvi

List of Figures xviii

Chapter 1 Introduction 1

1.1 The Case for Logical Sense-Making 2

1.2 Interdisciplinary Approach 4

1.3 Contributions and Outline of this Work 5

I Sense-Making 7

Chapter 2 Interpreting Pronominal Expressions 8

2.1 Introduction 8

2.2 Cognitive Science of Language and Anaphora 14

2.3 Interpreting Anaphoric Expressions in an NLP Pipeline 19

2.3.1 Non-Statistical Approaches 20

2.3.2 Statistical Approaches 21

2.4 From NLP Pipelines to Interactive Communication 22

2.4.1 Givenness 23

2.4.2 Common Ground 23

2.4.3 Consciousness and Cognitive Status . . . 25
2.4.4 From the Common Ground to the Common *Play*ground . . . 26
2.5 Interpreting Anaphora by Sense-Making - An Overview . . . 28
2.6 Contributions to Pronominal Anaphora Resolution . . . 30
2.7 Solving Pronominal Anaphora Problems in Imperative Discourse . . . 32
2.7.1 Overview of a Proof-of-Concept System . . . 32
2.7.2 Mathematical Preliminaries . . . 33
2.7.3 Detailed Walk-through of an Example . . . 35
2.8 General Properties . . . 44
2.8.1 Class of Situated Anaphora Resolution Problems . . . 44
2.8.2 Domain-Independent Aspects of the Reasoners . . . 45
2.9 Other Related Work . . . 46
2.9.1 Coreference Resolution . . . 46
2.9.2 Pronoun Disambiguation . . . 46
2.9.3 Reference Resolution in Robotics . . . 47
2.9.4 Natural Language Understanding in Robotics . . . 48
2.10 Some Limitations . . . 49
2.11 Conclusion . . . 51

Chapter 3 Interpreting Indirect Speech Acts 54

3.1 Introduction . . . 54
3.2 Approach . . . 56
3.2.1 Corpus Analysis . . . 57
3.2.2 Reasoner Development . . . 58
3.2.3 Related Work . . . 60
3.3 Reasoning for ISA Interpretation . . . 61
3.3.1 Preconditions for Interpretation . . . 62
3.3.2 Domain-General Reasoning . . . 63
3.3.3 Contextual Evidence . . . 68
3.4 Generating ISA Schemas - Building an ISA Corpus . . . 71

3.5 Some Related Work in Corpus Building for Knowledge-Based Agents 72
3.6 The "ISA Schema" . . . 73
3.7 Techniques for Developing a Corpus of ISA Schemas . . . 76
3.7.1 Expert Authoring . . . 76
3.7.2 Non-Expert Authoring . . . 77
3.7.3 Non-Expert Validation . . . 77
3.7.4 Defining a Task to Collect Data . . . 78
3.7.5 Extracting from Existing Corpora . . . 78
3.8 A General Approach to Developing a Corpus of ISA Schemas . . . 79
3.9 Example Development of a Corpus of ISA Schemas . . . 81
3.10 Discussion and Future Work . . . 87
3.11 Conclusion . . . 89

Chapter 4 Perceiving Object Affordances 91

4.1 Introduction . . . 91
4.2 Cognitive Science Background . . . 93
4.2.1 The Concept of Affordances . . . 93
4.2.2 Affordances in Cognitive Robotics . . . 95
4.3 The Computational Cognitive Affordance Framework . . . 98
4.3.1 Logic-Based Representation . . . 98
4.3.2 Computational Architecture (CALyX) - Overview . . . 99
4.4 Robot Kitchen Helper Experiment: Using and Handing Over Objects 103
4.4.1 Mathematical Preliminaries . . . 104
4.4.2 Semantic Representation of Visual Perception, F . . . 106
4.4.3 Relevant Contextual Items, C . . . 108
4.4.4 Cognitive Affordances, A . . . 110
4.4.5 Cognitive Affordance Rules, R . . . 112
4.4.6 Handover Inference – Introduction . . . 117
4.4.7 Inferring Affordances with Uncertain Logic . . . 117
4.4.8 DS-Theoretic Handover Inference . . . 119

4.5 Experiment - Multi-Domain, Multi-Scenario Handover 126
4.5.1 Introduction 126
4.5.2 Domains 127
4.5.3 Representing Social Cues and Domain Rules 129
4.5.4 Representing the Domain Distinctions 130
4.5.5 Experimental Scenarios 131
4.5.6 Experimental Results and Discussion 133
4.6 Discussion 135
4.6.1 Implications of the Proposed Architecture 135
4.6.2 Learning the Rules 139
4.7 Using Inferred Affordances - Future Work 139
4.7.1 Novel Tool Use 140
4.7.2 Creative Problem Solving 141
4.7.3 Role of Affordances in Sense-making 141
4.8 Conclusion 141

II Learning Knowledge – Case of Norms 143

Chapter 5 Learning Social Norms from Natural Language 144
5.1 Introduction 144
5.2 Theory of Cognitive Affordances 146
5.3 Grounding and Learning 148
5.3.1 Enabling Affordance Processing in a Cognitive Robotic Architecture 149
5.3.2 Learning Affordance Rules from Instruction 152
5.3.3 Executing Affordance-Based Commands 156
5.4 Evaluation 161
5.4.1 Simulation Experiment and Empirical Demonstration 161
5.4.2 Commands with Implicit Affordances 166
5.5 Discussion 167

5.6 Related Work . . . 169
5.7 Conclusions and Future Work . . . 170

Chapter 6 Cognitive Science of Norms Representations 172

6.1 Introduction and Motivation . . . 172
6.2 A Representation Format for Norms . . . 174
6.3 Norm Representation and Activation in Human Data . . . 176
6.3.1 Methodology . . . 176
6.3.2 Experimental Results . . . 179
6.4 Learning Norms . . . 181
6.4.1 How Do People Learn Norms? . . . 181
6.4.2 Data Representation Format of Norm Learning . . . 182
6.4.3 Algorithmic Learning of Experimental Data . . . 183
6.5 Context-Shifting . . . 186
6.6 Context-Specific, Belief-Theoretic Norm Representation . . . 186
6.6.1 Mathematical Formulation of a Norm System . . . 186
6.6.2 Evaluation: Dynamic Context Shifting . . . 188
6.7 Conclusion . . . 192

Chapter 7 Scaling up Norms 196

7.1 Introduction . . . 196
7.2 Representing Norms with DS-Theory . . . 198
7.2.1 Basic Notions in DS-Theory . . . 201
7.2.2 Relationship to Bayesian Theory . . . 202
7.2.3 Bundling Behaviors into Contexts with Indexed FoDs . . . 203
7.3 Agent Model of a Norm Learner . . . 205
7.3.1 Model of Evidence Received . . . 205
7.3.2 Learning Normative Prevalence $[\alpha, \beta]$ through Incremental Update . . . 206
7.3.3 Learning Normative Demands $\mathcal{D}$ from Sanction Signals . . . 208
7.4 Experimental Results . . . 209

7.4.1 Simulation Setup . . . 210
7.4.2 Experiment 1: Dynamics of Belief Update . . . 210
7.4.3 Experiment 2: Performance and Computational Complexity . 213
7.4.4 Experiment 3: Uncertainty due to Learner's Distance from Behavior . . . 216
7.4.5 Experiment 4: Uncertainty due to Occlusions in the Learner's Line of Sight . . . 217
7.5 General Discussions . . . 218
7.5.1 Importance of Sanctioning . . . 220
7.5.2 Assumptions about the Observations . . . 221
7.5.3 Managing Contexts . . . 222
7.5.4 Individual vs System-level Patterns . . . 222
7.5.5 Expressivity of Norms . . . 223
7.6 Related Work . . . 224
7.7 Conclusion . . . 226

Chapter 8 Consent **227**

8.1 Introduction . . . 227
8.2 Legal Landscape of Consent under the Law of Intentional Torts . . . 231
8.2.1 Conventions, Norms and Laws . . . 231
8.2.2 Actual Consent . . . 233
8.2.3 Apparent Consent . . . 235
8.2.4 Presumed Consent . . . 237
8.2.5 Constructive Consent . . . 237
8.2.6 Reluctant consent . . . 238
8.2.7 Other Issues . . . 239
8.3 Consent Issues in Robot Applications . . . 241
8.3.1 Robot Vacuum Cleaners . . . 241
8.3.2 Robot Waiters . . . 242
8.4 Research Directions for HRI . . . 244

8.4.1 Applicability of Consent . . . 245
8.4.2 Detecting Consent . . . 248
8.4.3 When Consent Does Not Matter . . . 251
8.4.4 Robot Roles . . . 253
8.5 Near-Term Next Steps for HRI . . . 255
8.5.1 Early-Stage Interaction Design . . . 255
8.5.2 Experimental Evaluation and User Studies . . . 256
8.5.3 Revisiting Past HRI Experiments . . . 258
8.6 Conclusion . . . 260

III Sense-Breaking 262

Chapter 9 Human Problem Solving 263

9.1 Introduction . . . 263
9.2 Problem Solving, Creativity and Insight . . . 266
9.2.1 What is Real World Problem-Solving? . . . 266
9.2.2 Analytical Problem-Solving . . . 267
9.2.3 Creativity . . . 270
9.2.4 Insight Problem Solving . . . 272
9.3 Event-Triggered Mode Switching During Problem-Solving . . . 275
9.3.1 Impasse . . . 275
9.3.2 Defocused Attention . . . 276
9.4 Role of the Environment . . . 276
9.4.1 Partial Cues trigger Relevant Memories through Context-Shifting 277
9.4.2 Heuristic Prototyping facilitates novel associations . . . 279
9.4.3 Making Physical Inferences to Acquire Novel Information . . . 280
9.5 Proposed Theory of Creative Problem Solving . . . 280
9.5.1 Dual Attentional Modes . . . 281
9.5.2 RWPS Model . . . 283
9.5.3 Model Predictions . . . 287

9.6 Experimental Challenges and Paradigms 290
9.7 Conclusion 292

Chapter 10 Formalizing the MacGyver Problem **294**

10.1 Introduction 294
10.2 The Turing Test and its Progeny 296
10.3 The MacGyver Framework 298
10.3.1 Formal Definition of a MacGyver Problem 299
10.3.2 Complexity Results and Importance of Heuristics 300
10.3.3 MacGyver Solution via Domain Modification 303
10.3.4 Connections to Insight Problem Solving in Humans 305
10.4 A Conceptual Blueprint for Solving Cup World 306
10.5 Agent Capabilities and Subtasks 309
10.6 Evaluating Agents and Measuring Research Progress 310
10.7 Conclusion 312

Chapter 11 Solving MacGyver Problems **314**

11.1 Introduction 314
11.2 The MacGyver Problem 316
11.2.1 Defining a MacGyver Problem 316
11.2.2 Solving MacGyver Problems 317
11.3 Proposed Algorithms 318
11.3.1 Overall Problem Escalation Procedure 318
11.3.2 Recognizing Anomalies 320
11.3.3 Modifying Action Definitions 322
11.3.4 Discovering New Actions 324
11.3.5 Generating New Constants 325
11.4 Robot Integration and Experiments 328
11.5 Other Related Work 329
11.6 Open Problems 330
11.7 Conclusion 332

Chapter 12 Conclusions **333**

12.1 Language Understanding 334

12.1.1 Using hidden assumptions for resolving pronouns 334

12.1.2 Using hidden assumptions for understanding speech acts . . . 335

12.1.3 Future work 335

12.2 Common Sense Assumptions about Social Norms 336

12.2.1 Social Norms Associated with Objects and Interactive Behaviors 336

12.2.2 Generalized Norm Representation and Learning 337

12.2.3 Future work 338

12.3 Breaking Assumptions: From Common Sense to Creativity 339

12.3.1 Formalizing creative problem solving 339

12.3.2 Future work 340

12.4 In Conclusion 340

Bibliography **341**

List of Tables

2.1 Computed uncertainties for entity candidates for pronominal anaphoric expressions in three situations. 53

3.1 Difficulty level of ISA schemas. 84
3.2 Results of Non-Expert Validation Study 1. 85
3.3 Results of Non-Expert Validation Study 2. 86

4.1 Domain-Specific Rule Uncertainty Assignment 131
4.2 Perceptual Uncertainty Assignment for 4 Scenarios 132

5.1 A subset of the relevant rules used by the natural language understanding component (NLU). 154
5.2 Ground Truth affordance rules. 163
5.3 High level syntax of understandable utterances, in JSpeech Grammar Format (JSGF). 164

6.1 Eliciting Probes for Three Norm Types 178
6.2 Origin of Selected Actions for *Library* Scene 179
6.3 Permission Norms for *Library* and *Jogging* Scenes in the Norm Generation Experiment 180
6.4 Uncertainty Intervals for Agents 1 and 2. 195

7.1 Parameter selections for each experiment in the simulation. 211
7.2 Experiment 2: Performance metrics over four forms of uncertainty. . 214
7.3 Experiment 2: Complexity Results 215

11.1 Experimental results for a Fetch robot attempting to solve a Cups World problem. 327

List of Figures

2.1 Example of commanding a robot to move blocks. 21
2.2 Example of incorrect resolution of pronouns by state-of-the-art neural coreference resolution systems 22
2.3 Architecture for resolving pronominal anaphoric expressions. 35

3.1 Types of reasoning required to interpret the utterance "Can you do (action) A?" 62
3.2 Example ISA Schema. 75
3.3 Suggested Context Features and Values, manually extracted from corpora by the authors 80
3.4 Example of the web-based GUI used for non-expert authoring of ISA Schemas. 80
3.5 Example of the web-based GUI used for non-expert validation of ISA Schemas. 84
3.6 ISA Schema authored from non-expert crowdsourcing. 87

4.1 Context-Sensitive Cognitive Affordance Model 98
4.2 Architecture for perceiving and reasoning about cognitive affordances. 120
4.3 Selecting Applicable Rules Based on Context 120
4.4 Determining Beliefs of Perceptual Aspects 121
4.5 Performing Inference of Cognitive Affordances 122
4.6 Results of affordance inference in various situations. 134

5.1 Cognitive Architecture for Learning Affordances. 150
5.2 Physical grasp affordances for a knife. 157
5.3 Affordance operations across the action hierarchy. 159

6.1 Four sample scene pictures used to elicit norms 177
6.2 Single run of the norm learning algorithm across two contexts. 193
6.3 Single run of norm learning, under contextual uncertainty. 194

7.1 Model for a Normative Agent 205
7.2 Experiment 1: Change in uncertainty over a simulation run under different experimental conditions of uncertainty. 212
7.3 Experiment 3: Comparison with Bayesian agent when there is uncertainty due to observational distance. 218
7.4 Experiment 4: Comparison with Bayesian agent when there is uncertainty due to observational occlusion. 219

9.1 Summary of neural activations during focused problem-solving and defocused problem-solving. 283
9.2 Proposed Model for Real World Problem Solving. 284

10.1 Example of a simple AI domain in which the agent must engage in creative problem solving. 296
10.2 Conceptual diagram showing several exemplary MacGyver problems and how they relate to classical planning tasks. 301

11.1 Problem solving when the agent must explore the subsymbolic space. 327

Chapter 1

Introduction

"Can you describe the shark?"

Source: The New Yorker Collection

"Can you describe the shark?" This was a caption from Harry Bliss' New Yorker cartoon in which a lifeguard on a beach is yelling the question to a distressed swimmer in the ocean. What would it take for an artificial intelligence (AI) agent to appreciate the humor in this cartoon? We humans effortlessly understand the absurdity involved here. Even though we do not see a shark, we assume that the swimmer is distressed because the shark is attacking them. In addition, we make an assumption about social norms being violated — namely, that the lifeguard should be displaying more urgency and rushing to save the swimmer rather than posing mundane questions from the comfort of their chair.

1.1 The Case for Logical Sense-Making

Making and breaking these kinds of assumptions is crucial to human intelligence and creativity. They fill gaps in our knowledge and resolve ambiguities in an otherwise endless stream of perceptual information. When broken, they allow us to restructure our knowledge in new ways and discover previously unknown connections. At the core of the process of making and breaking assumptions is the ability to continually assess and **make sense** of the world, resolving ambiguity to favor explanations that make the most sense, and recognizing anomalies as possible avenues for creative exploration.

Modern state-of-the-art AI systems do not "make sense" of data in this way. Instead, they learn by extracting statistical patterns present in the data. They learn to associate an input data sequence with its "most-likely" or statistically relevant mental abstraction. Often, these statistical estimates are learned in conjunction with a particular task like prediction or classification or clustering. So, sense-making for today's AI can be crudely summarized as the process of finding the most likely statistical explanation for the data for a particular task.

In this thesis, I argue that statistical sense-making alone is insufficient and that future AI systems will need to be able to do sophisticated logical reasoning over their own knowledge. For this, they will need different types of knowledge (causal, social, etc.) represented symbolically, and sophisticated reasoning machinery to perform logical sense-making operations. One way to clarify the overlapping roles of statistical and logical sense-making is to relegate logical and symbolic sense-making to higher-level cognition, i.e., the sort we engage with when we deliberate or think deeply about things. I show that such a clear division of labor is not realistic and sense-making (with reasoning) occurs at all levels of cognitive processing – from basic language understanding at the word-level, basic visual perception and action selection to high-level deliberation and even creative problem solving.

Beyond interpreting the humor in New Yorker cartoons, future AI and robotic systems will need sense-making abilities to both work effectively *with* humans and

to solve problems autonomously. Consider the following idealized, and arguably desirable, interaction between a human and their robot collaborator working together to fix a loose screw.

HUMAN: *Ah! This screw is loose. There should be a screwdriver in the toolkit. Can you grab it?*

ROBOT: *Okay.*

ROBOT: *[Returning with a coin] I didn't find one, but here is a coin that might work.*

HUMAN: *Isn't that Bob's coin?*

ROBOT: *Yes, but I checked with Bob and he said we could use it.*

(1)

First, in order to be helpful, the robot would need to *understand* the human request. The robot has to be able to decipher word meanings, including words that do not carry any semantics like the word "it". At the sentence level, the robot would need to understand that the question "Can you grab it" is not really a question to be answered verbally, but a request to perform some actions. In addition, the robot has to be able to handle the presumably novel scenario of a missing screwdriver and improvise with another object (a coin in this case). In doing so, it must reason about the affordances needed for the task, and potentially try different objects, and maybe even different coin sizes. Visual stimuli alone might be insufficient to determine the viability of the coin. Social norms like that of the ownership of the coin might need to be considered. One coin might make more sense than another if it has all the right properties and appropriate normative restrictions have been lifted. The central claim in this dissertation is that if future AI systems and robots have to be able to not only engage in such a dialog but also actually perform the underlying tasks then

they must be able to *reason* over abstract conceptual knowledge about world.

1.2 Interdisciplinary Approach

Immanuel Kant, the 18th century German philospher argued that sense-making is the process of assimilating sensory information into a coherent, unified whole along with other ideas [Kan08]. Weick elaborated on this definition, stating that sense-making is the process of structuring the known by placing stimuli into some kind of framework that enables us to comprehend, understand, explain, attribute, extrapolate and predict [Wei85].

But, what exactly is this "unified whole" or "framework"? What does it contain? How exactly can it assimilate stimuli? I explore these questions, looking for properties of mental representations and computational operations needed to make sense of a variety of sensory stimuli – from language to visual perceptions. My approach is interdisciplinary in nature: I explore a variety of AI domains, formalize different types of knowledge and reasoning needed within these domains, develop algorithms and cognitive architectures for reasoning with this knowledge. Developing effective approaches requires bringing together many disciplines covering areas of knowledge representation and reasoning, machine learning, natural language processing, uncertainty processing, human-robot interaction, as well as cognitive science, social psychology and ethics.

Building AI systems requires a number of representational, algorithmic and architectural commitments. One unifying theme in this thesis is that despite the variety of representational forms used in this thesis, they are largely symbolic, sometimes with numeric uncertainty. Representations are central to sense-making and they allow an agent to abstract over their world. While this thesis argues for symbolic representations and first order logic, it does not commit to any one fragment or instantiation in particular. Sense-making for language understanding uses a dialect of logic programming as a symbolic representation, while work on affordances uses uncertain logic. The work in creativity uses state transition models that are

represented via classical planning. Each chapter covers a certain domain of AI and accordingly will feature its own representational notation and characteristics.

1.3 Contributions and Outline of this Work

The plan of this thesis is as follows: This chapter introduced the key issues that I address, discussing what sense-making is and why it is needed and what sorts of computational requirements it demands. The major challenge of the thesis is in showing exactly how sense-making, a somewhat nebulous notion, can be computationalized.

Part I takes up this challenge, discussing the role of sense-making in understanding language and in action selection. Specifically, Chapter 2 focuses on one specific language interpretation task – pronominal anaphora resolution – in which I discuss the sorts of knowledge and inference machinery that might be needed for resolving ambiguity that arises when interpreting pronouns like "it." Chapter 3 extends the approach of Chapter 2 to another language interpretation task – indirect speech act resolution – that is at a higher-level of processing (namely sentence level) as compared to pronoun resolution, which deals with words.

Chapter 4 demonstrates the role of sense-making when using and manipulating objects – perceiving and actualizing object affordances. Specifically, I discuss how reasoning under uncertainty can be carried out in an integrated goal-oriented robotic system that comprises different and distinct components tasked with processing visual percepts, understanding dialogue and sequencing actions commensurate with commands and tasks. An interesting question that is raised in chapters 2, 3 and 4 are how can some of the knowledge used for sense-making be learned?

In **Part II**, we take a deep dive into one specific kind of knowledge that is often used in sense-making – social norms. Social norms are a form of commonsense knowledge that are particularly elusive and difficult to computationalize, and lack training data. Chapter 5 begins by outlining how norms associated with object manipulation, discussed in Chapter 4, can be learned from Natural Language Instruction.

Chapter 6 considers how humans learn norms from observation and provides a computational framework for how these norms can be learned automatically under uncertainty. Chapter 7 extends the norm-learning model in Chapter 6 by scaling it, while tackling challenges with complexity and uncertainty.

Although elusive, this normative knowledge is important to sense-making and pervasive in human reasoning and interaction. Chapter 8 explores one specific set of norms – "consent" – and considers how legal theory can help navigate these norms and be incorporated into human-robot interaction research.

In **Part III**, we flip the discussion towards sense-breaking, which in a sense is complementary to sense-making, and is the basis for creative problem solving. Specifically, Chapter 9 focuses on human creative problem solving and dives deep into its neuroscientific and psychological underpinnings. Chapter 10 uses the models hypothesized in Chapter 9 and proposes a computational formalism, dubbed "MacGyver Problems," for novelty discovery, anomaly detection through a sequence of sense-breaking and sense-making mechanisms. In Chapter 11, I demonstrate a set of algorithms for solving MacGyver problems.

Chapter 12 concludes the thesis.

Part I

Sense-Making

Chapter 2

Interpreting Pronominal Expressions

Part I tackles how AI systems can better understand language and select how objects can be used with logical reasoning over explicitly represented symbolic concepts.

In this chapter, I focus on the role played by logical sense-making in deciphering word meaning, specifically looking at pronominal anaphoric referring expressions (e.g., "it") and identifying what entity these expressions refer to in imperative (Action-oriented) discourse.

2.1 Introduction

> *"If an incendiary bomb drops near you, don't lose your head. Put* ***it*** *in a bucket and cover* ***it*** *with sand."* [Hir81].

Anaphors are linguistic referring expressions whose interpretation depends on objects and entities introduced earlier in a discourse [Mit14]. For example, in the sentence: "pick up the parcel and give it to me," the pronoun "it" is an anaphor that relies on its antecedent "the parcel" for its meaning; both mentions likely pointing to the same real-world parcel. Pronouns (and anaphors more generally) are used extensively in dialogue and discourse and there are often multiple antecedent candidates

for an anaphor, leading to ambiguity, a problem that humans handle quite gracefully [GS86]. Anaphora resolution is a central problem in natural language understanding. Here, I focus on a subclass of this problem involving object pronouns when they are used in simple imperative sentences (e.g., "pick *it* up.").

But, how do we, humans, use "it" in imperative discourse? Consider the following discourse between A and B in a scene that consists of a floor on which there is a parcel:

A:	*Pick up the parcel.*
B:	*Okay. [B picks up the parcel]*
A:	*Open it*
B:	*Okay. [B opens the parcel]*

(2)

In order for B to follow instructions – first to pick up the parcel and then to open it – B must interpret the definite referring expressions "the parcel" (a definite noun phrase) and "it" an unstressed third-person pronoun (which we will henceforth reference in bolded italics as ***it***). This means B must identify the real-world object that A is referring to in the discourse. Now, let's say B is a robot that is equipped with a "parcel detection" algorithm and therefore is capable of translating the lexical content of the expression "the parcel" and is able to identify the single object on the table that matches this linguistic description. However, what about ***it***? The pronoun ***it*** has no lexical or semantic content of its own, and is purely anaphoric, that is it's meaning is tied to a previous mention. It should be plainly obvious that ***it*** in dialogue (2) refers to the same parcel. More specifically, ***it*** and "the parcel" are co-referent with the same discourse referent. In the above discourse situation, resolving ***it*** is quite trivial. However, things can get hairy quite quickly, as we will see below:

[Tabletop with a parcel and a mug]

A: *Pick up the parcel that's next to the mug.*

B: *Okay. [B picks up the parcel]*

A: *Open it.*

B: *Okay. [B opens the parcel]*

(3)

In dialogue (3), there are two discourse referents (i.e., real world objects) and two definite expressions – the parcel and the mug. Again, if B had a parcel detector and a mug detector, it will be able to resolve these expressions. Resolving ***it*** however is more complicated as there are multiple possible antecedent candidates. Which object does ***it*** refer to? We know that the answer is probably the parcel, and we might justify so by saying that mugs cannot be opened. [1] But, what if the affordances offered by both objects match the issued imperative?

[Tabletop with a parcel and a mug]

A: *There is a parcel on the table that I want.*

B: *Okay.*

A: *Walk over to the table. Carefully move the mug out of the way. Then, grab it.*

(4)

In dialogue (4), ***it*** could refer to the parcel, the mug or the table. All three objects can technically be grabbed. However, we know that ***it*** refers to the parcel, even though the parcel was not mentioned in the previous utterance or even the one

[1] We might also argue that the parcel is more salient as it is the object of an imperative sentence (where the subject is implicit), whereas the mug is simply part of the prepositional phrase.

before. It seems that the parcel was the most salient item, and we might be able to interpret the discourse as a sequence of instruction for grabbing the parcel, allowing B to keep its eyes on the prize, namely the parcel. In some sense, A's communicative intent was to request B to pick up the parcel. But, can ***it*** refer to items not the central or most salient item in the discourse?

[Scene: Tabletop with a parcel and a mug]

A: *[pointing to the mug] I want that mug. Pick up the mug and put it in the parcel* (5)

B: *Okay. [B puts the mug inside the parcel]*

A: *Bring it to me.*

In (5), we have two ***it*** mentions. First, occurs at the first utterance where ***it*** refers to the mug as there is only one antecedent. However, the second ***it*** occurs in the *bring* imperative and has a slightly different antecedent. Here, ***it*** refers to both the parcel and the mug (or the mug inside the parcel). B cannot simply assume that ***it*** only refers to the mug because doing so would result in B removing the mug that it just placed in the parcel. Moreover, there is no antecedent linguistic mention of the parcel-mug combo, which makes resolving ***it*** more difficult as the pronoun itself has no semantic content to help with interpretation. This missing antecedent notion can be further complicated, as follows:

A: *Open the front door. There is a surprise waiting for you.*

B: *Okay. [B opens the front door, and notices a parcel on the mat. A sees B noticing the parcel and B knows that A is watching him.]*

A: *Go ahead and bring it in.*

(6)

Once again, in (6), ***it*** refers to the parcel and there is no linguistic mention of any parcel.

To summarize, the unstressed third-person pronoun ***it***, although seemingly simple, does not inherently contain any semantic or lexical content of its own and derives its meaning either explicitly from antecedents or implicitly from other modalities. What makes ***it*** resolution so interesting and challenging as a phenomenon to theorize about is the variety of ways in which it can be used. Here we reviewed a number of considerations that must be accounted for in a theory that explains the production and comprehension of referring expressions such as ***it***. This is only the tip of the iceberg of the different ways ***it*** is used, which can additionally include pro-sentential (e.g., "Donald Trump did not tweet today. ***It*** caused a panic on Fox and Friends"), pro-verbal usage (e.g., "stop doing ***it***!"), and non-referential (e.g., "***It*** is raining."), to name a few. An ***it***-Theory must be able to account for having to (1) select one from multiple antecedent candidates, (2) find the correct antecedent at a position earlier in discourse before the most recent mentions, (3) select antecedents that could be from different syntactic categories, (4) resolve references without explicit linguistic antecedents, and (5) handle discourses comprising a plurality of ***it***'s each potentially tied to a different discourse referent.

While the examples presented above are helpful to understand the challenges in resolving ***it***, they are not exactly how humans speak to each other. We do not

speak in full, clear, perfectly-paced, optimally-packaged sentences, but instead we produce fragments called "turn construction units" (TCUs), and interact in a conversational "ping-pong" game[2] organized as adjacency-pairs [dRng]. Moreover, there are numerous umms, pauses, stresses, overlaps, repairs and interruptions that make the conversation (if transcribed more carefully) look less like the examples above and more like this one below[3]:

```
33 *H: A:sk him if the REP is out.
34     (0.8)
35 *CC:Is the REP out? (0.6)
36     (0.5)
37 *C: That's affi::rmati::ve,
38     (31.9)
39 *P: Carnarvon, Gemini-5, the preliminary look, it's still drifting out a
40     little bit, looks like it got about 5.8 feet..
```

First, we can see that some of same challenges noted above when resolving ***it*** also appear in this real human communication data. Clearly, the parties are saying much more than what is being captured in my earlier parcel-examples, and this additional information could be crucial for resolving anaphoric expressions as they might suggest topical and salience shifts. In this dissertation, I do not explore these more naturalistic issues, choosing instead to simplify and abstract the question of what sorts of cognitive mechanisms might be needed to resolve referring expressions. This would be a view that is decidedly outside of the purview of Conversation Analysis (CA) [HK17], a rich research tradition which places more emphasis on the data itself and pushes for a more inductive inquiry driven by raw, but fully annotated, natural conversations. My focus in this thesis is on the cognitive mechanisms underlying comprehension and production of referring expressions, which in some sense is more aligned with psycholinguistics and computational linguistics and less so with CA. Nevertheless, I hope that my discussion will not have precluded any

[2] A metaphor that emerged from discussions with J.P. de Ruiter.

[3] Many thanks to Saul Albert for this transcription of Gemini-V air-to-ground communication involving several parties in Canarvon, Houston and the space module.

future possibility for an inductively driven CA inquiry.

2.2 Cognitive Science of Language and Anaphora

The current view of anaphora resolution is that it depends on context. Anaphoric referring expressions are interpreted with respect to a discourse model which is built up dynamically while processing a discourse and which includes objects that have been mentioned (Mitkov 2014). This simple idea is actually quite profound and suggests (1) multiple levels of representation, where there may be (2) top-down and bottom-up influences between the levels. Moreover, processing is likely to be (3) incremental, and is driven by (4) satisfying constraints to resolve temporary (5) ambiguities that arise as a result of partial sentence comprehension. There is a separate question about whether current theories about anaphora are (6) modular or interactive theories and whether the processes proposed operate in (7) parallel or serial. The examples I provided above are imperatives issued by a speaker to a listener that the latter has to follow. Thus, (8) language in imperatives is not merely a product but is an action or a speech act, which raises questions about the intentions of the speaker in these examples. Another crucial question is when (i.e., time-course) the anaphor is resolved – is it (9) predicted? In all these aspects, there is an underlying question of how best to choose (10) the appropriate task. Finally, there is a question of (11) how words and concepts – lexicon – are stored. Below, I will present a glossary of sorts introducing and describing each of the above eleven questions in the context of existing literature in word recognition, syntactic and semantic parsing and perspective-taking. These various fundamental concepts in modeling language comprehension apply with equal force to anaphora resolution, as well.

Levels of Representation Language comprehension models usually describe processing within and/or between different "levels" of representations. The term "levels" means different things to different people and here are a few interpretations. First, a level can represent the type of knowledge or a knowledge system.

In this interpretation, each level can be thought of as a system that comes packaged with its own unique way of describing and representing knowledge, and its own set of rules about this type of knowledge. As Jackendoff (2002) suggests, there might be structure within a particular level, i.e., embedded internal levels (intra-level). Second, a level can represent the continuum between perceptual (concrete) and cognitive (abstract) knowledge. For example, a concrete perceivable acoustic signal for a word might be at a different level from the abstract word or concept represented by that signal. Third, a level can represent a construct that is a combination of other levels. That is, while two representations might be at one level, their combination might be at a different one. Fourth, levels can be viewed against cognitive psychological notions of executive and cognitive control, goals, and motivations. In the word recognition literature we can see two examples: Elman and McClelland's (1988) interactive activation model and Marsen-Wilson's (1987) Cohort model. In the E&M model, the levels of representation were lexical, phonemics and acoustics. In M-W, the levels were messaging (or higher-level considerations), lexical, and phonemic.

Top-Down and Bottom-Up Processing The notion of levels is necessary to discuss top-down and bottom-up effects. In general, top-down implies an influence of a higher level on a lower level. When modeling coarticulatory influences, Elman and McClelland (1988) considered how knowledge of word would influence the perception of a phoneme – this is a top down effect as the lexical layer was consider a higher level than the phonemic layer. The impact of acoustic information on the phonemic layer, in this model can be considered a bottom-up effect. As I will discuss later, these effects do not have to be symmetric.

Incrementality One important question in sentence comprehension is the timing of when the sentence is processed – is it processed at the end after all the bottom up information is received or is there some continual incremental processing? By incremental processing I mean that all sources of info are processed as quickly as they become available and the information that has been processed is then used to process new upcoming information. In many of the models in word recognition as well as sentence processing, higher-level representations are constructed and updated as

each word is received. The notions of incrementality, levels, and top-down processing set the stage to discuss an important class of language processing models: constraint-satisfaction.

Constraint-Satisfaction Many language models view comprehension as a constraint-satisfaction problem. Consider the Cohort model in which the goal of the task was to recognize a word (e.g., "chair") as quickly as possible. As one hears the word, but before the uniqueness point, there is simply not enough bottom up information to recognize the word. Yet, we recognize such words quickly and before the uniqueness point. One explanation is that multiple types of knowledge (or levels) serve as constraints that can assist in comprehension. None of the constraints by themselves are sufficient to uniquely identify a word. However, what is particularly powerful is the relationship between constraint-satisfaction and incrementality in the resolution of ambiguity, which I will discuss next.

Resolving Ambiguity with Constraint-Satisfaction during Incremental Processing The concept of incrementality suggests that information is processed in real-time as it is received. Crucially, that information is made available to whomever (e.g., other levels) needs it, immediately. Higher-level representations (e.g., message level, discourse models) are built up on the fly as new information arrives. For example, when processing sentences on a word by word basis, as each word is received, assigning a syntactic parse or semantic mapping or even other functional mappings like resolving anaphors is made difficult by the existence of temporary ambiguity because the entire sentence (or discourse) is yet to be received. Constraints help with this temporary ambiguity problem because the knowledge in the sources of constraints are made available immediately and can be used to incrementally process the input (MacDonald, Pearlmutter, & Seidenberg 1994). Thus, this constraint-knowledge immediately helps language disambiguation and comprehension in real-time.

Modularity versus Interactivity and Parallel versus Serial In language comprehension, questions of representations are closely intertwined with questions of process because processes compute or in some way work on representations (Gar-

nham 2010). I have briefly touched on the computational aspects of "process" in discussions about constraint-processing and incrementality. However, I have not yet discussed the structural properties of these sorts of systems. The first concept in this regard is the notion of modularity of a cognitive process, which captures ideas of encapsulation, independence, and lack of interactivity with other processes. Modularity cannot be discussed without first defining levels and representations and types of knowledge. On the opposite end of modularity is interactivity, where levels of representation influence each other, and sometimes even influence the inner workings of the other (full interactivity). In word recognition research, E&M model (1988) can be considered to be interactive in the sense of symmetry as the different levels influenced recognition (both top down and bottom up directions). On the other hand M-W's (1987) Cohort model placed the bottom-up processing from the phonemic level in a privileged position (bottom-up primacy), thus only being partially interactive. Interactivity does not necessarily mean that the observed effects need to happen immediately; we can have delayed effects in both modular and interactive theories as E&M demonstrated in their paper. A crucial point about modularity is that if two processes A and B are said to be modular, then process A is not influenced by the output of process B, even if B's output is available. We also saw notions of modularity in dual route systems where one handled most cases and another handled exceptions. Modularity is an orthogonal notion to whether processes are parallel or serial. The Frazier and Rayner model for handling garden path sentences was a modular theory that was also serial. On the other hand, MP&S (MacDonald, Pearlmutter, & Seidenberg 1994) and TT&G (Trueswell, Tanenhaus & Garnsey 1994) propose interactive models that are also parallel. Modular processes can occur in parallel if they co-occur, but their results may be merged at some later point. M-W's Cohort model can be thought of a model that is partially interactive and serial as there is a particular ordering of processes on the levels, but these processes can influence each other. In general, the notion of parallel can suggest that two processes are occurring at the same time. It is also possible to think about parallel in terms of an intra-level phenomena in which multiple, say syntactic, parses are possible and

processed in parallel.

Language as Action Language is used for communication, none more obvious than the anaphora examples presented earlier in the context of imperative dialogue. Not only is the sentence or speech act derived from a communicative intent held by a speaker, but these communications are occurring in the real world with referents available all around the interlocutors. This view is different from language comprehension work discussed earlier and consequently necessitates different experimental paradigms and models. Crucially, the language comprehension models must be able to incorporate the speaker. For example, in the literature on resolving referential expression and perspective-taking, the visual world paradigm was used that allowed for modeling a comprehender that assumes something about the speaker in a shared space of ideas (common ground) (Tanenhaus, Chambers & Hanna 2004). This sort of view is particularly relevant for interactive communication and anaphoric reference resolution, and is exemplified by Goodman's Rational Speech Act Theory (RSA), which captures this essence in Bayesian models (Goodman 2016).

Prediction There are two ways to think about prediction. One view is as a "builder" or "helper" in which prediction helps recognition of component parts. Language processing does some work ahead of time (before receiving bottom up input) and this work helps facilitate recognition. The alternative view is to consider the role of prediction as that of a scientist. The mind is in the business of explaining, so it forms hypothesis and then collects evidence (via percepts) to test and update these hypotheses. In this view, the goal is to minimize prediction error. Comprehension is then the idea of interpreting a sentence to identify a linguistic signal given some generative model. In Goodman's RSA model, the comprehender uses a generative model for a speaker and the speaker uses one for the listener in a recursive probabilistic programming paradigm.

Role of Task In all these above-mentioned concepts, there is an underlying question of how best to choose the appropriate experimental task. A key component to consider is the idea of "goals." For example, in M-W the goal was to recognize a word. Other goals can include "arrive at a single interpretation of a sentence," or

"disambiguate different antecedent candidates", or "parse words into syntactic positions," or more generally, "determine what the speaker is saying." The task imposes a goal to satisfy constraints. The relevant constraints that are activated then influence how the stimuli are processed. Thus, the results of any language comprehension experiment have to be evaluated against the backdrop of what goals are imposed by the task, which in turn have to be considered against what constraints are evoked.

Representing Concepts in a Lexicon The lexicon usually refers to the stored knowledge about words. That said, this is a loaded concept and it is still unclear what representations the lexicon contains and how lexical access is accomplished. The traditional view is one of a mental dictionary with rules and exceptions. In this view, discrete entities corresponding to a word are typically accessed. A more updated and modern view is that the lexicon is more like an index that provides pointers or references to the various other memory stores of knowledge (syntactic information, meaning, etc.) where the particular concept can be found.

2.3 Interpreting Anaphoric Expressions in an NLP Pipeline

Research into computational approaches for resolving anaphoric expressions (particularly pronouns) within the field of Natural Language Processing (NLP) is decades old and quite extensive. Early computational approaches were largely non-statistical and worked with syntactic and semantic parses of sentences to work out the meaning of pronouns. More recently, as statistical machine learning and data-driven approaches started gaining popularity, computational models have begun treating anaphora resolution (pronoun disambiguation) as equivalent to co-reference resolution[4], which in turn is handled by discovering patterns in the co-occurrence between the pronoun and various antecedent candidates.

Interpreting pronouns has traditionally been thought to involve satisfying a set of "constraints," and if several antecedent candidates still remain, then choosing a preferred one based on various "preference" factors [Mit14]. Typical constraints

[4]This approximation, in my view, is quite problematic as it is entirely disembodied from the context in which the utterance and discourse are situated.

that have to be satisfied are: (1) morphological constraints (e.g., gender, number and person agreement), (2) syntactic constraints (e.g., "John likes him" vs "John likes himself"), and (3) semantic constraints (e.g., the infelicitous use of the pronoun in "John doesn't have a car. *It is in the garage."). In addition to these constraints, certain interpretations are generally considered dispreferred, a quality that can be useful when we have to decide between multiple antecedent candidates that satisfy all three constraints. Preferences are related to the use of commonsense knowledge, the restrictions on type of argument a verb may possess, syntactic considerations that prefer object pronouns with object antecedents, and finally salience [Mit14]. Salience is a particularly important one and is a widely exploited preference factor. It has many different forms, two of which we discuss here are recency and repetition. Referents introduced in recent utterances are more likely to be referred to by a pronoun than those introduced in utterances further back (recency). Referents that have already been referred to frequently are more likely to be pronominalized than those that have not (repetition). The various computation approaches avail themselves of these constraints and preferences, and in this section, I will discuss some non-statistical and statistical approaches that make computational commitments to some (but not all) of these constraints and preference factors.

2.3.1 Non-Statistical Approaches

Hobbs presented one of the earliest algorithms for pronoun resolution in [Hob78], which exploits the syntactical and morphological constraints by searching syntactic trees of current and preceding sentences in breath-first and left-to-right manner, stopping when it found a matching noun phrase [Hob78]. This was largely considered naive and is generally unable to handle ***it*** resolution when there are multiple candidates such as those in my example earlier. Hobbs' subsequent "semantic" approach (complementing his syntactic approach) consideres the role of word meaning and suggests that making logical inferences over an extended set of concepts in knowledgebases may be helpful in disambiguating pronouns.

Lappin and Leass introduced in [LL94] a more sophisticated algorithm that

Figure 2.1: Consider instructing a robot the following: *"Pick up block1. Put it on block2. Pick up block3. Put **it** on block1."* Which block does the second "it" refer to? This question is trivial for humans, but not so for many artificial systems.

maintains and updates a discourse model that tracks the salience of various mentioned referents based on recency and grammatical function. It is able to discover mentions up to four sentences back. However, these early approaches do not reason over other aspects of the discourse that could potentially be relevant. They are unable to handle the examples (such as (4)) presented earlier.

2.3.2 Statistical Approaches

With the gaining popularity of statistical machine learning methods, some researchers began using Bayesian methods to compute the probability $\mathrm{p}(a \mid p, f_1 \ldots f_n)$, where a is an antecedent candidate of a pronoun p and f_i are features obtained from constraints and preferences [GHC98]. This method uses the Hobbs algorithm to identify all possible antecedents and then uses a large corpus (Penn Treebank) to compute likelihoods needed. More contemporary approaches aim to solve the sister-problem of co-reference resolution using statistical and neural network based algorithms. Interested readers can find nice overviews of the evolution of this field in Mitkov ([Mit14]) and Ng ([Ng10]). The state-of-the-art systems are trained on large corpora and they are able to recognize mention-pairs that are statistically related to one another.

However, many of these systems fail when presented with the simplest of examples. Consider the simple example of the commanding a robot to move blocks shown in Figure 1. This is trivial for us to resolve. But, as shown in Figure 2, Clark and Manning's system in [CM16], a leading deep neural network-based system, does not produce the correct resolution of this example.

Extracting statistical patterns is insufficient because disambiguating pronouns can require inference on world models that evolve through the discourse.

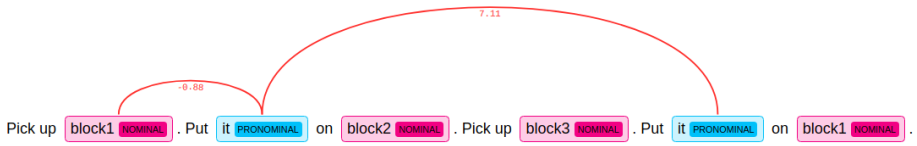

Figure 2.2: Incorrect resolution of a current neural coreference system [CM16] on the example shown in Figure 2.1.

Thus, it is unsurprising that many current co-reference systems do not resolve this simple example because it is not the statistical relationships that undergirds disambiguation, but instead extra-linguistic knowledge. Some have proposed specific representations to incorporate background knowledge needed to solve these more "hard co-reference problems" [PKR15]. However, these systems still lack the ability to reason about the discourse itself, which I will argue is necessary.

2.4 From NLP Pipelines to Interactive Communication

The computational approaches presented thus far treat natural language understanding as a modular and linear process, solely from either the speaker's or listener's/reader's perspective, but not both. These computational models all subscribe to Shannon and Weaver's [SW63] model for communication. In this model, information flows from the speaker to the listener unidirectionally, and comprehension occurs by decoding speech signals received at the listener. Pronoun resolution approaches that rely on this model assume that disambiguation at the listener does not need to consider the speaker, but just the properties inherently present in the utterance data itself. Human communication, however, is not unidirectional and clean in this manner. Instead, it is a highly dynamic and bilateral set of behaviors that can influence the other and alter subsequent responses [NDRA10]. I now turn to some approaches that directly incorporate this crucial joint speaker-listener aspect of communication into their computational models.

2.4.1 Givenness

First, it might be helpful to consider the notion of "givenness," which can be thought of as the cognitive informational status associated with a discourse referent along the dimension of whether they are new or previously "given." The level or degree of givenness then tracks how these known items may vary along this dimension. Referring expressions are used by the speaker not for new information but for items already given. This generative (or speech production) view of givenness can be one way to begin unpacking what referring expressions refer to what objects. Baumann and Riester have argued in [BR12] that there are two ways one can look at givenness: (1) as whether an item is given in the common ground, and (2) as whether an item is in the listener's consciousness.[5] We will consider each view, in turn, below.

2.4.2 Common Ground

Stalnaker defines common ground as common knowledge that is mutually recognized as such by the speaker and listener [Sta02]. It is generally understood that definite referring expressions are based on information that is in the common ground [CM02]. The relevance of common ground on language production and comprehension has been a subject of active debate (the perspective-taking debate), with some arguing for the primacy of common ground [CM02, BSH18] and others arguing that speakers produce referring expressions more from their own ego-centric perspective [Key07] (see [BSH18] for a nice overview of this debate). While this debate is not entirely settled, it is now generally agreed that common ground does play a role (at least to some degree) in reference resolution. So what does the common ground contain?

A recent computational model by Goodman and Frank in [GF16] takes a Bayesian approach for speech act comprehension and production that recursively considers speakers and listeners in a probabilistic programming paradigm. In their proposed Rational Speech Act theory (RSA) there are three interactive models - one for a literal listener who interprets semantics of an utterance literally, one for a ratio-

[5] There is a third way, that I will leave out as it does not directly add to our discussion for object pronouns.

nal speaker who considers how a literal listener might interpret a referring expression (based on its discriminatory strength), and one for a pragmatic listener who considers this speaker model when interpreting the referring expression. RSA has shown to map onto human performance quite well in generating and interpreting definite references. However, the highly contrived and artificial experimental setting severely limits the findings and consequently the ecological validity of this approach. In the case of pronouns, RSA is especially handicapped as there is no semantic content in the pronoun for the literal listener to work with, which means that the rational speaker has no utility function that it can use. In the case of definite noun phrases, the rational speaker was able to use a utility function that captured the discriminatory property of a reference. Without lexico-semantic information in ***it***, it becomes difficult to construct useable utility functions for the rational speaker. Without a suitable utility function, the pragmatic listener has no likelihood information about the speaker model.

An even larger challenge with RSA is that it is a computational level account [MP76] and provides no guidance (beyond a superficial utility function) about the types of information in the common ground or that are needed for resolving references. That said, the use of Bayesian updates is a more powerful way to model human-like communication than the Shannon and Weaver approaches espoused in the previous section. de Ruiter and Cummins propose such a general model of human communication in [DRC12] in which communicative intent is identified online during the relevant utterance using prior knowledge along with new information. They propose the AIRBUS model for inferring communicative intention rapidly using precomputed likelihood database. While promising, the model does not specify exactly how it can be adapted into a language processing model that resolves pronouns.

Thus, while the notion of common ground is quite compelling as a way to view and analyze givenness, current approaches have not directly addressed how (from a process point of view) this common ground can be leveraged for pronoun resolution.

2.4.3 Consciousness and Cognitive Status

The other way to view givenness of an entity is by considering whether the listener has that discourse referent in their consciousness at the time of hearing the utterance. Gundel et al. in [GHZ93] proposed a framework for associating forms of referential expressions with a presumed cognitive status of a listener. The framework comprises of six hierarchically nested tiers of cognitive status. Each tier is cued by a set of linguistic forms, for example the "in focus" or "focus of attention" (FoA) tier is cued by the pronoun ***it***.

As noted by Williams in [WS18], the GH framework interfaces nicely with general pragmatic principles operative in language interpretation, such as Grice's Maxims [Gri75]. Thus, GH can facilitate interpretation in tasks such as the reference resolution task. However, it is not a process model, so it cannot directly produce a solution to a reference resolution problem. Williams adopted the GH-theoretic framework into such an AI process model (computational model that can actually perform a language processing task). Williams proposed an algorithm for navigating the GH-Theoretic hierarchy to produce both the appropriate reference for a speaker (generation) and the appropriate discourse entity (resolution).

Williams as well as others ([CPQ06, Keh00]) have employed the GH-Theoretic framework to handle pronouns such as ***it*** by referring to discourse entities in the tier called "Focus of Attention." Crucially, they apply the anaphoric salience preference of recency for populating the FoA. The most recent noun phrase in the previous sentence is usually selected and included. As I have outlined, recency alone is not sufficient in handling many other varieties of pronoun usage. There are many cases where the most recent definite noun phrase is the incorrect one (examples (2), (3), and (4)).

Thus, while these approaches show promise in that they provide more tangible process models for resolving referring expressions, they are still lacking for a couple of reasons: (1) their accounting for how discourse entities populate the cognitive status groups is somewhat limited (e.g., recent noun phrases are placed in the focus

of attention), and (2) there is limited accounting of common ground in that it is still unclear how knowledge in the common ground is utilized in language processing. When is the common ground used, what does it contain and how is the information in it employed?

2.4.4 From the Common Ground to the Common *Play*ground

Stalnaker noted in [Sta02] that all conversational moves are performed in relation to the common ground. Common ground has traditionally been thought to contain representations of information that is physically co-present, linguistically co-present, as well as information about cultural or community membership that is likely to be jointly known [CM02].

The perspective-taking debate largely centers around conflict between theoretical findings in [CM02] and empirical findings within a constrained visual world paradigm (see [Key07]), and around conflicts even within these empirical findings. The questions being addressed in these debates are centered around when, during language processing, do speakers and listeners consider common ground and when they consider their own egocentric (or privileged) ground. I do not intend to enter into this debate in this essay, but instead nudge the discussions about common ground away from questions of relevance and timing to ones of membership and nature: what is actually contained in the common ground and how is it represented and used?

First, consider the question of whether or not a piece of information is in the common ground (membership). Brown-Schmidt suggests that [BSH18] prior research adopted a limited, binary view of membership: discourse referents are either in the common ground or not. Per Brown-Schmidt, recent research argues otherwise noting that representations can be quite rich, suggesting that representation of entities in the common ground are not binary but graded and probabilistic.

Second, with regards to what sorts of information is contained in the common ground (nature) and how it can be used, Clark and Marshall note in [CM02] that interlocutors maintain rich representations of their daily interactions in episodic

memory. These representations, which form the basis of inferences about common ground, include information about what happened, what was said, what was seen, and by whom these events were experienced. However, in the name of experimental control and tractability, empirical research adopted a simpler propositional stance towards the common ground, limited to named discourse entities or some superficial property (e.g., color, facial feature, height or shape). So, what are these richer representations contemplated by Clark and Marshall?

I propose that the nature of common ground is that of a collection of logical reasoners - that is the speaker assumes not only that the listener has access to a certain set of facts or propositions about the discourse referents, but also assumes that the listener has access to certain sets of inference mechanisms and situation-independent inference rules, and therefore the ability to compute implicit facts, determine satisfiability and entailment. This means that the information structure used to resolve definite references look less like a passive common ground, and more like an active and vibrant playground. The common playground is full of interesting equipment (reasoners, inference mechanisms and rules) that can transform beliefs in real-time as the discourse unfolds. Due to its dynamic nature, the common playground allows the interlocutors to run simulations and see if antecedent candidates make sense (i.e., are satisfiable). External cues and bases for mutual understanding (e.g.,eye gaze and pointing, physical co-presence) can provide clues as to what aspects of the common playground are considered salient within the listeners mind and therefore speak to the cognitive status of discourse referents. Specifically, a referent might receive probabilistic determinations of its membership status to the focus-of-attention-tier based on whether it "makes sense," when placed in the common playground. Stalnaker foresaw the possibility of this more active notion in [Sta02] when he suggested that the common ground should account for rule-like inferences and "salience of consequence," but cautioning that "even if I see that something follows, I may not be sure that you will or that you think that I will or ..."

In the next section, I present, with the help of an example, a brief sketch of a potential model of pronoun resolution that incorporates this notion of a com-

mon playground into a GH-theoretic setting, thereby unifying the two notions of givenness noted earlier. In subsequent sections, I show how this model can be computationalized and implemented in an artificial system.

2.5 Interpreting Anaphora by Sense-Making - An Overview

Language comprehension can be viewed as an incremental model-building and generative process (see [Kam81]) in which the listener must either perform (or at least mentally simulate) the issued actions thereby changing the surround world (or simulated world). In doing so, resolving anaphors becomes a task of associating actions with their parameters in a way that "makes sense" in this unfolding narrative. Listener's resolution process is guided by the common playground, which in turn is influenced by the listener's own capabilities, expectations of the speaker (and if relevant, other interactants), and the general normative climate and community expectations. I propose that listeners, when tasked with disambiguating ***it*** as part of an imperative, must bind the issued action to one of a set of two or more antecedent candidates, and then ask three sense-making questions: (1) *can* the listener perform the desired action on an antecedent candidate, (2) *should* they perform the desired action on the candidate, and (3) is the speaker *intending* for them to perform the desired action on the candidate. Each of these three questions in turn spawn separate reasoning mechanisms that can then attempt to answer them.

An example walk-through of the resolution process might make things a bit more clear. Consider a scenario where A is cooking and B is A's sous chef. In this scenario, we can consider one of two situations: (1) A is washing dishes or (2) A is mixing a salad. In both these situations, B is responsible for cutting some tomatoes. In this scenario, A says to B, ""Pick up the knife. Cut the tomato." Once B has cut the tomato, A says "pass it to me." It becomes immediately clear that ***it*** resolves to the knife in situation (1) and to the cut up pile of tomatoes in (2)[6]. How would B

[6]One could argue that this is not felicitous as the chopped pile is really a plurality of tomato pieces and A would say "pass them to me." But, this is a debatable point as it is possible that A could say "pass it to me" as they are considering ingredients as a whole mental concepts.

resolve this?

We can consider three reasoners (one for each sense-making question noted earlier): (1) an affordance reasoner for evaluating the *can* question, (2) a normative reasoner for evaluating the *should* question and (3) a speaker intent reasoner for evaluating the *intent* question. When activated, the three reasoners come pre-loaded with some *situation-independent* knowledge. For example, the affordance reasoner comes with some rules about effect of actions, inertia axioms and some common sense physics laws about the actions relevant to the discourse (e.g., objects cannot be in two locations at once, one cannot pick up something one is already holding, etc.). The normative reasoner might come pre-loaded with rules about normative functions of objects (e.g., knives used for cutting, tools can be washed, edible items can be incorporated into a meal, etc.). Speaker intent reasoner might come pre-loaded with rules about communicative intent and preferences of speakers (e.g., objects are more relevant to a speaker if they will help a speaker, and speakers do not intend for listeners to pass them irrelevant things).

During processing, common *situation-specific* facts about the object candidates and scene information is provided into each of the reasoners. Prior to receiving the situation-specific facts, the reasoners are incomplete and cannot be run. However, once situation-specific facts arrive, the reasoner can generate satisfiability results. Using appropriate non-monotonic reasoning methods, we can compute, for each object candidate, the likelihood that its inclusion into the reasoner will result in satisfiability. We can then probabilistically combine these results from the reasoners as pieces of evidence that can be used to arrive at probability distribution over the discourse entities.

While the situation-independent knowledge (rules and facts) might be different for each reasoner, what is common across all the reasoners are the situation-specific facts. What allows us to tie the results from all the reasoners together is that they are all reasoning about an action-object relationship, i.e., the relationship between action (issued in the imperative) and objects (antecedent candidates). I implemented such a system using the language of Answer Set Programming, details

of which follow.

2.6 Contributions to Pronominal Anaphora Resolution

Consider these three examples (including the one presented in the previous section), all situated in a kitchen where a robot is assisting a human with a cooking task:

A: *Pick up the [knife]$_k$. Cut the [tomato]$_t$. Put* ***it***$_{k,t}$ *down.* (7)

B: *Okay.*

A: *Pick up the [knife]$_k$. Cut the [tomato]$_t$. Put* ***it***$_{k,t,b}$ *in the [bowl]$_b$.* (8)

B: *Okay.*

[Speaker context - A:Washing dishes, B: Cooking]

A: *Pick up the [knife]$_k$. Cut the [tomato]$_t$. Pass* ***it***$_{k,t}$ *to me.* (9)

B: *Okay.*

In each of these cases, the disambiguation of "it" requires one to consider not just the statistical relationships (like those inferred by the coreference systems) or static and timeless bits of commonsense knowledge (like those used in tackling WSs), but contextual information available to an agent that is situated and embodied in an environment. It is unclear what types of knowledge or reasoning capabilities are needed. In the most general case, the problem is very hard and been the subject of research for decades [Hob78, LL94, Win80]. However, in the narrower case of imperative dialogue we can simplify the problem by focusing on the cause and effect relationships associated with *performing* actions issued by the speaker.

The goal of this chapter is to unpack this problem of situated or embodied anaphora resolution. We focus on object pronouns (like "it") as used in imperative utterances within a larger discourse. We view natural language comprehension as an incremental model-building and generative process [Kam81] in which the listener must either perform or simulate the issued actions thereby changing the surrounding world. In doing so, resolving anaphors becomes a task of associating actions with its parameters in a way that "makes sense" in this unfolding narrative. Specifically, the contributions are:

1. **(Problem Characterization)** We introduce the general class of situated anaphor resolution problems in imperative discourse. We characterize these problems by providing a set of exemplary problems and some insights into what makes them particularly special and distinct.

2. **(Proof of Concept)** We construct a proof of concept system using Answer Set Programming and Dempster-Shafer theory for solving this class of problems. The system can resolve the ambiguous anaphors in 7, 8 and 9. We present a detailed walk-through for (5).

3. **(Reasoning Characterization)** We articulate some general and domain-independent types of reasoning as well as architectural capabilities needed to solve these problems.

2.7 Solving Pronominal Anaphora Problems in Imperative Discourse

2.7.1 Overview of a Proof-of-Concept System

To solve situated anaphora problems, a listener agent must reason about extra-linguistic information obtained as a result of its embodiment (i.e., sensory-motor and bodily capabilities of the agent) and its situatedness (agent interactions in context with its environment, which includes other agents). The agent's decision making is guided by the mutual knowledge shared with its interactants, which in turn is influenced by the agent's own capabilities, expectations of its interactants and general normative expectations of the society in which the agent is situated [CM02].

We propose that when tasked with disambiguating an anaphoric object pronoun as part of an imperative (e.g., "pass $\text{it}_{\text{knife,tomato}}$ to me."), the agent must bind an **action** to one of a set of two or more **object candidates**. To do so, it must reason about *three different aspects* over and above syntactic considerations, which together form the mutual knowledge, namely:

1. **Plausibility**: *Can* it perform the desired action on an object candidate?
2. **Normative**: *Should* it perform the desired action on an object candidate?
3. **Speaker Intent**: Is the speaker *intending* for it to perform the desired action on an object candidate?

We formalize these notions by suggesting that these three aspects or reasoning modes can be structured as microtheories and represented as answer set programs. Reasoning within these microtheories can happen in parallel with each reasoner returning uncertainty measures for each object candidate. We propose then combining uncertain evidence obtained from these theories using Belief-theoretic notions of evidence combination. Belief theory (a subset of which is Demspter-Shafer theory) generalizes Bayesian probability theory and provides some unique advantages over Bayesian updates to modeling epistemic and subjective uncertainty. Moreover, it

has a rich history of application in sensor-fusion networks, which the proposed proof of concept system is modeled after.

In the next section, we walk through a demonstrative example in more detail. But, first, we provide some background on Answer Set Programming and Dempster-Shafer theory and provide some intuition for why they might be suitable frameworks for resolving situated anaphors.

2.7.2 Mathematical Preliminaries

2.7.2.1 Answer Set Programming.

Answer Set Programming (ASP) is a knowledge representation language useful for commonsense reasoning, especially in presence of incomplete information, defaults, exceptions and inductive definitions [Bar03a]. A logic program Π is a set of rules of the form:

$$L_0|\ldots|L_k \leftarrow L_{k+1},\ldots,L_m, \textbf{not } L_{m+1},\ldots,\textbf{not } L_n$$

Where L_is are literals in the sense of classical logic and the **not** represents *negation-as-failure*. The left and right hand sides of the rule are called the *head* and *body* of the rule, respectively. Either one of head and body can be empty. When the head is empty, i.e., $k = 0$ and the $L_0 = \bot$ the rule is called an *integrity constraint*. When the body is empty, the rule is called a *fact*. Intuitively the above rule means that if $L_{k+1},\ldots,L_m$ are true and if there is not proof that $L_{m+1},\ldots,L_n$ are true (i.e., can be safely assumed to be false), then one of $L_0|\ldots|L_k$ must be true. The semantics of ASP is based on the stable model semantics of logic programming [GL90b].

ASP serves as a suitable language with which to represent knowledge in the proposed microtheories for solving situated anaphora resolution problems, for several reasons. First, ASP allows *non-monotonic reasoning*, that is adding more knowledge can change one's previous beliefs, a mode especially true of situated reasoning when the world state and context can change and evolve. Second, because ASP allows

for negation-as-failure (**not** L_i) and classical negation ($\text{-}L_i$), default rules can be encoded, which as we will see, allows for encoding complicated cases where, for example, certain actions are not permissible if there is no reason to think they are not forbidden. Third, ASP allows for what are known as *choice rules*. In addition to literals, the head of the rule can contain *cardinality constraints* of the form $l\{L_0, \ldots, L_k\}u$ in which l, u are integers and explicitly allow the encoding of choices. Finally, we will need to be able capture dynamic systems when reasoning about actions and ASP, through its implementation as an *incremental logic program* which allows for capturing knowledge accumulating over increasing time steps [GKK+08]. For this chapter, we use ASP implementations in `clingo` and `iclingo`, which provide both grounding and solving capabilities.

The language of answer set programming is very expressive allowing us, under brave and cautious reasoning, to express every property of finite structures that is decidable in the complexity classes Σ_2^P and Π_2^P, respectively [EGM97, FL07].

2.7.2.2 Dempster-Shafer Theory.

DS-Theory is a measure-theoretic mathematical framework that allows for combining pieces of uncertain evidential information to produce degrees of belief for the various events of interest [Sha76] . In DS-Theory a set of elementary events of interest is called *Frame of Discernment* (FoD). The FoD is a finite set of mutually exclusive events $\Theta = \{\theta_1, ..., \theta_N\}$. The power set of Θ is denoted by $2^\Theta = \{A : A \subseteq \Theta\}$. Each set $A \subseteq \Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping $m_\Theta(\cdot) : 2^\Theta \to [0, 1]$ such that $\sum_{A\subseteq\Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set $\mathcal{F}_\Theta$. The triple $\mathcal{E} = \{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$ is called the *Body of Evidence* (BoE). For ease of reading, we sometimes omit $\mathcal{F}_\Theta$ when referencing the BoE. Given a BoE $\{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B\subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) =$

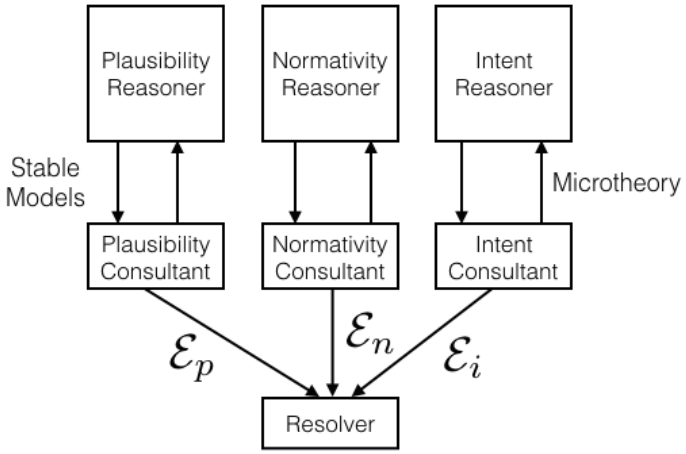

Figure 2.3: Approach for resolving situated anaphors. Each of the three proposed reasoning modes are encoded as microtheory templates, filled with situational information by a consultant and solved using a reasoner (e.g., answer set solver). Uncertainty measures are computed over the set of object candidates and combined to return the best guess.

$1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict A. The *uncertainty* interval of A is $[Bel(A), Pl(A)]$, which contains the true probability $P(A)$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$.

DS-Theory extends Bayesian theory in several ways, allowing for some capabilities that are suitable for our purposes. First, it allows for assigning probabilistic measures to sets of these hypotheses (not just individual ones), including the set of all hypothesis. This allows DS-Theory to consider missing and ambiguous information, which is helpful when there is evidence that an anaphor could resolve to more than one object candidate. Second, DS-Theory does not require assuming any prior distributions over object candidates, which is useful when priors are difficult to justify. Bayesian and DS-theories do share many commonalities and DS-theory is often viewed as being a generalization.

2.7.3 Detailed Walk-through of an Example

Consider the discourse D_1 from (5). In the scene, the speaker is performing a [washing dishes/cooking] task and is looking to have the robot use a knife to cut the tomato and then to pass over the [knife/tomato] once it is done. Below is the dis-

course comprising three utterances, as follows:

Pick up the $[\text{knife}]_{\text{k}}$.

Cut the $[\text{tomato}]_{\text{t}}$.

Pass $\text{it}_{\text{k,t}}$ to me

We propose a knowledge representation scheme and resolution approach implemented through a resolver architecture as shown in Figure 2.3. Specifically, we consider three consultants, one for each reasoning mode. The consultants instantiate and run a reasoner, and later compute uncertainty metrics for object candidates. Each reasoner is an ASP solver that operates on a microtheory which is a grounded logic program. The consultants maintain partial or incomplete microtheories (*microtheory template*) that are domain-independent. During resolution, the consultants "fill-in" the missing facts and rules in order to be able to complete reasoning. The microtheory templates are not solvable logic programs as they are not sufficiently ground, but merely serve as a general blueprint for a consultant to flesh out. In the next section, we will discuss properties of these templates in more detail, but for now, we will look at specific microtheories for our running example.

Once reasoning has been performed, each reasoner returns, to its corresponding consultant, answer sets (if available) that are stable models consistent with the information that the agent has used in reasoning. Thus, each reasoner answers the question of whether a set of facts relating to the discourse "makes sense" from a plausibility or normativity or intent point of view, depending on the consultant. The consultant defines a DS-theoretic mass function over these models, which in turn allows the consultant to build a BoE ($\mathcal{E}$) over the objects of interest. We will now look at each microtheory for (5) in more detail.

2.7.3.1 Plausibility Microtheory.

The plausibility microtheory contains knowledge for determining if an *action* requested in an utterance *can* be performed in relation to an *object*, given the agent's action capabilities and the current situation. We use an incremental answer set pro-

gramming paradigm as it is reasoning with action dynamics. First, we establish a program base that incorporates a set of facts true of the agent's (initial) current situation in its interaction with a human (named "commX" or "commander X")

```
% For incremental mode iclingo
#include <incmode>.

#program base.
% Percept types
is(obj1,object).
is(obj2,object).
is(obj3,scene).
is(obj4,loc).
is(obj5,person).
is(self,loc).

% Percept names
has(obj1,name,knife).
has(obj2,name,tomato).
has(obj3,name,kitchen).
has(obj4,name,table).
has(obj5,name,commX).

% Fluents
init(has(obj1,loc,self)).
init(has(obj2,loc,table)).

% Initial state axiom
holds(F,0):- init(F).

% Basic action definitions
action(pickup(X)) :- is(X,object).
action(putdown(X)) :- is(X,object).
action(pass(X,Y)) :- is(X,object),is(Y,person).
```

The next step is to GENERATE a set of action occurrences using ASP choice rule syntax, as follows:

```
#program step(t).
%GENERATE
{ occ(A,t) :  action(A) } = 1.
```

The syntax states that at any given time instance 't', one and only one action occurs from the set of possible actions. We can then define the effects of our actions and various action-related axioms.

```
%DEFINE
% Effect of action occuring
holds(has(X,loc,self),t) :- occ(pickup(X),t-1).
-holds(has(X,loc,table),t) :- occ(pickup(X),t-1).
holds(has(X,loc,table),t) :- occ(putdown(X),t-1).
-holds(has(X,loc,self),t) :- occ(putdown(X),t-1).
holds(has(X,loc,Y),t) :- occ(pass(X,Y),t-1).
-holds(has(X,loc,self),t) :- occ(pass(X,Y),t-1).

% Inertia axioms
holds(F,t) :- holds(F,t-1), not -holds(F,t).
-holds(F,t) :- -holds(F,t-1), not holds(F,t).

% Commonsense laws
% Objects cannot be in two locations at once.
-holds(has(X,loc,Y),t) :- holds(has(X,loc,Z),t),is(Y,loc),
Y!=Z.
```

Once, the axioms are defined we use the integrity constraint syntax of ASP to represent various action requirements and various situations that do not make plausible sense.

```
%TEST
% Cannot pick up something you are already holding
:- occ(pickup(X),t), holds(has(X,loc,self),t).
```

```
% Cannot pick up something when holding something else
:- occ(pickup(X),t), holds(has(Y,loc,self),t).

% Cannot put down something you are not holding
:- occ(putdown(X),t), -holds(has(X,loc,self),t).

% Cannot pass something you are not holding (on the table)
:- occ(pass(X,Y),t), holds(has(X,loc,table),t).

% Cannot pass if recipient already has it
:- occ(pass(X,Y),t), holds(has(X,loc,Y),t).

#program check(t).
:- query(t), t<maxlength.

#const maxlength=5.
```

The `#program check(t)` operation provides a termination of the program, which in this case is after five steps, set by the `maxlength` constant. This microtheory is essentially a planning microtheory defined over a short time horizon of five steps.

2.7.3.2 Normative Microtheory.

The normative microtheory contains knowledge for answering the question of whether an action *should* be performed on an object candidate. This microtheory shares many common features with the plausibility microtheory including the types of fluents and static predicates that are used such as $is(X, Y)$ and $has(X, Z, Y)$. These predicates are intentionally quite general and are designed to be representative of a high level language that a situated agent can use [BLS17]. A crucial difference between the normative microtheory and the plausibility microtheory is the choice rule in the "GENERATE" part, and more specifically, the special predicate ($has(A, permissible, X)$) that is used therein, as shown below

```
%GENERATE
has(obj5,is_doing,(washing_dishes;cooking)) = 1.

{ has(A,permissible,X) : is(A,action_verb),is(X,object),
is(S,person),has(S,uttered,A) } = 1.
```

The normative microtheory also has an additional choice rule associated with the task context of `washing_dishes` versus `cooking`. The task context choice rule is meant to allow the program to consider what would happen in different contexts. However, we anticipate that in real-world situations, the context is set and not necessarily choose-able in this manner. Nevertheless, we provide this as a choice rule, so we can compare performance across two different contexts. We can then define and test "permissible" actions against what is deemed forbidden or in some instances, what is not shown to be not forbidden, depending on the normative requirements at play.

```
%DEFINE
% Forbidden to pass something that is not
% a dish to someone who is washing dishes.
has(X,function,tool) :- has(X,used_for,cutting).
has(cooking,requires,X):- has(X,used_for,eating).
has(washing_dishes,requires,X):- has(X,function,tool).
-has(A,forbidden,X) :- has(S,is_doing,T),
has(A,permissible,X), has(T,requires,X).

%TEST
:- has(A,permissible,X),not -has(A,forbidden,X).
```

2.7.3.3 Speaker Intent Microtheory.

The speaker intent microtheory contains knowledge for answering the question of whether the action on an object candidate was what the speaker intended for the

agent to do. Once again, this microtheory shares many of the same predicates with the normative and the plausibility microtheories. And, as with those other two theories, what is unique is the special predicate used in the generate step ($has(A, speaker\_intends, X)$).

```
%GENERATE
%Task that the speaker is doing
has(obj5,is_doing,(washing_dishes;cooking)) = 1.

{has(A,speaker_intends,X) : is(A,action_verb),is(S,person),
has(S,uttered,A),is(X,object)} = 1.
```

The potential set of action-object pairings suggested by the speaker's utterance are constrained by what actions might be relevant to the task that the speaker is performing.

```
%DEFINE
has(X,nextAction,A) :- has(X,loc,self),is(X,object),
has(A,name, putdown).
has(X,nextAction,A) :- has(X,loc,table),is(X,object),
has(A,name,pickup).

% When something is split, it is made up of multiple parts
% Speaker unlikely to use "it" when referring to
% multiple object
has(X,number_parts,multiple) :-
has(X,physical_integrity,split),is(X,object).
-has(A,speaker_intends,X):-
has(X,number_parts,multiple),
has(A,verb_pronoun_ref,W),
is(W,pronoun),has(W,name,it),is(X,object),
not has(A,speaker_intends,X).

% speaker prefers the robot to perform the next action
-has(A,speaker_intends,X):- not
has(X,nextAction,A),is(A,action_verb),is(X,object), not
```

```
has(A,speaker_intends,X).


% relevance of an object to a speaker if it helps the
% speaker
has(X,function,tool) :- has(X,used_for,cutting).
has(cooking,requires,X):- has(X,used_for,eating).
has(washing_dishes,requires,X):- has(X,function,tool).
has(X,relevant_to,S) :- has(S,is_doing,T),
has(T,requires,X), is(S,person).

%TEST
% speaker does not intend for the robot to pass
% it irrelevant things.
:- has(A,speaker_intends,X), not has(X,relevant_to,S),
has(S,uttered,A).
```

2.7.3.4 Combining Evidence with Dempster-Shafer Theory.

We are reasoning about whether or not certain action-object pairings make sense. Thus, if the set of candidate objects is $O = \{o_1, \ldots, o_n\}$, we can define the DS-theoretic frame of discernment $\Theta = O$ (e.g., $\Theta = \{obj1, obj2\}$). Now, each microtheory potentially outputs a set of answer sets, $\mathcal{A} = \{A_1, \ldots, A_n\}$, where if $\mathcal{A} = \emptyset$ then the microtheory is unsatisfiable. We know that each answer set contains a set of ground predicates $A_i = \{p_i^1, \ldots, p_i^k\}$ including exactly one special generative predicate p_i^*. A generative predicate p_i^* is a predicate that is applied via choice rules. The special generative predicate is a domain general aspect of a reasoner that allows an agent to link an object with an action in a way that conforms to the needs of the particular reasoner. For each of our three aforementioned reasoners, we have one special generative predicate. For the plausibility reasoner, $p_i^* = occ(a(o), t)$, where a is an action, and o is an object from O (e.g., $occ(pickup(obj1), 1)$). For the normative reasoner $p_i^* = has(a, permissible, o)$ and for the speaker intent reasoner $p_i^* = has(a, speaker_intends, o)$. We can say that a grounded special generative predicate *contains* an object $o \in O$ when o is a term (or a nested term) in the pred-

icate. Each grounded special generative predicate *contains* one of the n candidate objects from O. Since the special generative predicates are introduced via choice rules requiring them in a reasoner, each answer set A_i will have a p_i^*, unless $\mathcal{A} = \emptyset$.

We can now define our mass function of some subset $B \subseteq \Theta, B \neq \emptyset$ as being the following:

$$m_\Theta(B) = \begin{cases} \frac{N}{|\mathcal{A}'|} & \forall b \in B, \exists i, \text{ such that } p_i^* \in A_i \text{ and} \\ & p_i^* \textit{ contains } b, \text{ does not } \textit{contain } b' \notin B \\ 0 & \textit{otherwise} \end{cases}$$

Where $\mathcal{A}' \subseteq \mathcal{A}$ refers to those answer sets that contain the desired action verb, and N is the number of answer sets $A_i \in \mathcal{A}'$ that satisfy the specified criterion. The reasoning consultants compute these mass functions and return a BoE $\mathcal{E}$ that contains the computed mass functions for the focal elements For example, if there are two object candidates $\Theta = \{o_1, o_2\}$ and a reasoner returns three answer sets $\mathcal{A} = \{A_1, A_2, A_3\}$, with A_1 containing o_1, A_2 containing o_2 and A_3 containing both object candidates, then the BoE will contain masses $m(\{o_1\}) = 1/3, m(\{o_2\}) = 1/3, m(\{o_1, o_2\}) = 1/3$. The evidence from these sources can be combined using the Dempster's rule of combination, which aggregates evidences or confidence values from different sources, but within the same frame of discernment. The results from computing the fused uncertainties for each of discourse examples (3), (4) and (5) are shown in Table 2.1.

2.8 General Properties

In this section, we discuss several general, domain-independent aspects of the proposed approach.

2.8.1 Class of Situated Anaphora Resolution Problems

Thus far, we have presented a few examples of situated anaphora resolution problems.[7] Discourses in this class of problems share the following features:

1. Each discourse consists of a set of utterances, at least one utterance being an imperative. At least one imperative utterance contains at least one anaphoric referring expression.

2. The discourse context contains, among other things, two or more candidate antecedents to the at least one anaphoric expression. The antecedents could be explicitly mentioned linguistic referring expressions or discourse entities (real-world objects or cognitive concepts) being considered by the interlocutors as part of the discourse.

3. Linking the anaphoric expression to one of the candidate antecedents requires reasoning with *extra-linguistic situational knowledge*.

4. Executing an imperative (or mentally simulating it) during the discourse can change the state of the world in a way that influences the interpretation of a subsequent anaphor. That is, the interpretation of the anaphor is not merely influenced by the semantics of a linguistic expression (as in WS), but by what happens in the world as a result of incrementally interpreting (and executing) each utterance in the discourse.

[7] Additional examples: (1) "Walk to the green door. Enter your passcode on the panel to open it." (2) "My pen fell under the bed. Grab that broom. Use the long end to get it out." (3) "The knife is on the chair. Pull it out. Grab the knife. Push it back."

2.8.2 Domain-Independent Aspects of the Reasoners

Each of the microtheories share a common structure as allowed by the GENERATE-DEFINE-TEST methodology in ASP. In the English language, every simple imperative utterance with an object pronoun has no overt subject (i.e., the subject could be, as in 'I would like you to ...') and the verb is often in its bare form. This focused structure allows us to consider specifically the relationship between just the *action verb* and the *object*. In the case of pronoun use, the object is replaced by an object pronoun such as "it." From a representational standpoint we are only interested in this relationship between *action* and *object*. This means we can pre-define useful relationships between action and object for each of the reasoners and generally ask the question if an action-object pairing makes sense from the corresponding reasoner's (or microtheory's) perspective. If we let a be the action specified in the utterance, and O represent the variable corresponding to the object pronoun, then we have three general symbols, one for each reasoning mode, as follows:

1. Plausibility Reasoning: $\mathbf{occ}(a(O), t)$
2. Normative Reasoning: $has(a, \mathbf{permissible}, O)$
3. Speaker Intent: $has(a, \mathbf{speaker\_intends}, O)$

Moreover, many of the commonsense definitions shown in the code fragments in the previous section for each of the reasoners are in fact domain-independent such as the axioms of inertia in the plausibility reasoner, the rule associated with when an action is not forbidden ($\text{-}has(A, forbidden, O)$) in the normative reasoner and the rule relating to when an object is relevant to a speaker ($has(O, relevant\_to, S)$) in the speaker intent reasoner. In fact, we were able to model (3) and (4) in much the same way as we did with (5).

We note that not only does this generality hold within a domain (like cutting vegetables), it extends to other domains as well as in cases where it is possible to incorporate several object pronouns in sequence, each referring to different objects, as follows:

Pick up the [ladle]$_l$. Put it$_l$ in the [pot]$_p$ containing [soup]$_s$. Stir it$_{l,p,s}$. Check if it$_{l,p,s}$ is mixed. Take it$_{l,p,s}$ out and wash it$_{l,p,s}$. (1)

2.9 Other Related Work

2.9.1 Coreference Resolution

Contemporary approaches aim to solve the sister-problem of coreference resolution using statistical and neural network based algorithms. Mitkov ([Mit14]) and Ng ([Ng10]) provide nice overviews of the evolution of this field. The state-of-the-art systems are trained on large corpora and they are able to recognize mention-pairs that are statistically related to one another. However, many of these systems fail when presented with situated anaphors, much like (1). Clark and Manning ([CM16])[8] propose one such system that does not produce correct resolution of (1) (see Figure 2.2), (3), (4), (5) and all the examples provided in the chapter. Extracting statistical patterns is insufficient because disambiguating situated anaphors requires inference on world models that evolve through the discourse. Some have proposed specific representations to incorporate background knowledge needed to solve these more "hard coreference problems" [PKR15]. However, these systems still lack the ability to reason about plausibility, normativity and speaker intent, which we argue to be important reasoning modes in situated cases.

2.9.2 Pronoun Disambiguation

There have been parallel efforts in tackling the Winograd Schema Challenge and the leading approaches employ a strategy of first selecting a format to represent commonsense factual knowledge, learning vast amounts of static commonsense knowledge from online databases (ConceptNet, WordNet and CauseCom), and then performing inference or classification with this knowledge [BHL+15, SVAB15b, LJL+16, GCS17]. However, these approaches do not consider the three reasoning modes we discuss.

[8] https://huggingface.co/coref/ and http://corenlp.run/

Also, it is unclear how the knowledge needed for performing this situated reasoning can be acquired from these online databases.

In another line of work, Richard-Bollans et al. have suggested that speaker intent, models of the situation, and choosing relevant entities are important considerations that must be taken into account when resolving pronoun disambiguation problems [RBGAC17]. They go on to propose the use of prototypes, plausible object configurations in the scenario described, and generally deeper semantic knowledge bases. While they do not provide a concrete proposal for addressing pronoun disambiguation using these principles, we agree with the basic premise that the type of knowledge needed for resolving references in the real world are more than merely formalizing knowledge from WWW datasets. We extend these ideas and suggest that the reasoning needed must incorporate deeper notions of context, plausibility, and normative standards.

2.9.3 Reference Resolution in Robotics

There has been considerable work in *reference resolution*, more generally, and in applying various theories from cognitive science in order to use pragmatics [Sch14, RBGAC17, Keh00, CPQ06, WASS16, WS16, VD16]. Unfortunately, much of this work focuses on pragmatics as it pertains to processing effort and cognitive effect on the agent, and less so on situational aspects of the agent's surroundings. The models that do unpack discourse context information have received a weaker computational treatment.

When embodied in the real world, an agent must be able to discuss, reason about and even perform actions on entities referenced in dialog, aspects that are somewhat unique to situational anaphora resolution problems described in this chapter. Accordingly, the agent must be able to create representations for those entities and incorporate them into the reasoning process. This process is more generally termed *reference resolution* and there has been extensive research in addressing how an agent might perform such resolution in the open world under conditions of uncertainty [Keh00, CPQ06]. Recently, Williams et al. have proposed a computa-

tional framework with which to resolve references to specific real-world co-present objects by formalizing a linguistic framework called the Giveness Hierarchy (GH) [WASS16, WS16] . However, the GH approach primarily performs syntactic disambiguation of pronouns (especially "it") using notions of recency of usage. Thus, it can resolve the references in certain WS correctly for example in (1), but not those from (2), (3) or (4). Nevertheless, one key strength of this approach is that an agent can represent a real world entity or percept abstractly, setting the stage for us to do more advanced reasoning capabilities over these abstracted entities. For example, a real world visual point cloud might be perceivable through an agent's visual sensors. This approach allows an agent to systematically tie this point cloud to an abstract entity (e.g., the symbol "obj1"), but also link this abstraction to other perceptual entities like the linguistic term "cup" and a haptic percept "heavy" and a cognitive concept "my cup".

2.9.4 Natural Language Understanding in Robotics

More generally, there is a vibrant research community exploring natural language communication with robots. Tellex et al. provide a extensive and current review of the current state of the art in both language understanding and language generation [TGKGM20].[9] Despite tremendous progress, missing from current work is importance of tracking the evolution of the discourse model as the language unfolds,

[9] An underlying assumption in much of this work is that language can only be understood when it is grounded in the perceptual aspects of the world, an idea originally proposed by Harnad [Har90]. The argument, roughly, is that symbols and mental concepts must be "grounded" in perceptual experience of the agent. The question of "symbol grounding" and whether or not symbols need to be grounded is an open philosophical question. Sloman has argued that the notion of symbol grounding is merely a rehashing of the philosophical idea of "concept empiricism," the idea that all concepts must have their basis, directly or indirectly in experience [SKCT06, Slo11]. Sloman reminds us that Kant refuted this notion by noting that certain concepts (e.g., space, time, ordering, causation) are needed to even make sense of experience. Sloman argues that certain concept can be (partially) defined implicitly by their role in powerful theories. Here, I do not take a position in this debate, but instead offer a cautionary slippery slope warning that a hard-nosed symbol grounding view runs the risk of turning natural language understanding into a linear pipeline of tasks from experiential percepts of text and audio into some representation of meaning, later to be used by the robot to plan and act upon. Such a linear pipeline goes against decades of research in how humans use language in dialogue and conversation, and how the process is not linear, but interactive, and even seemingly upstream tasks require downstream operations – e.g., interpreting ambiguous pronouns sometimes require reasoning about the actions and plans being talked about in an utterance.

particularly as it pertains to resolving references. Moreover, there is little to no work on the impact of social norms on language understanding. As I have shown, humans use norms and a range of implicit knowledge about actions and speaker intentions to make sense of language, even at the reference resolution level.

As alluded to earlier, the relationship between low-level perception and higher-level reasoning might be tight, and possibly symbiotic, so resolving ambiguities across levels might need to occur jointly. Recent work by Thomason et al. [TPS+20] proposes a dialog agent with some interactive capabilities and point in a promising direction. In this work, the dialog agent learns meanings of words, including words novel to the robot, through interactive multi-modal dialogue. The work provides a technique for jointly learning semantic representations and perceptual groundings for new concepts.

We are not aware of any prior approach that has been shown to be capable of reasoning with extra-linguistic information (including normative knowledge) about objects, when such information may be acquired in a situated manner from multiple sensory capabilities of an embodied agent. Moreover, this extra-linguistic information is typically mutual knowledge that is shared by the interlocutors, and includes knowledge about the agent's own capabilities, knowledge about other co-located agents and general extant normative or community standards.

2.10 Some Limitations

There are still many open questions with this model, one of which is the selection of a suitable model for speech production. The production model flips the script and requires: (1) selecting an action-object combination, (2) selecting the set of reasoners that the listener is likely to use, (3) for each reasoner selecting the set of situation-independent rules that the listener is likely to use, (4) selecting the set of potential antecedent candidates that the listener is likely to consider, and (5) running these reasoners, producing probability distributions over the antecedent candidates, and finally selecting the pronoun based on these distributions.

Another question is how does this approach fit into the prior discussion about common playgrounds and the GH-theoretic process models. We know from GH-theory that objects in the FoA (focus of attention) can be felicitously referred to by the pronoun ***it***. I suggest that the reasoners provide a mechanism by which this FoA can be populated, one that is more sophisticated than the use of the notion of recency in current approaches. The FoA could contain just one item (e.g., a MAP estimate from the probability distribution over possible antecedent candidates) or could itself be a probability distribution. The reasoning architecture (i.e., the common playground) then allows the speaker and the listener to populate their respective FoAs. That said, it is still a bit unclear how reasoning can work alongside other preference factors (e.g., salience) and constraints (syntactic and morphological) that can influence anaphora resolution.

One key question is whether this approach is computationally tractable and cognitively feasible. It appears that reasoning in this manner for each and every ***it*** is cumbersome and arguably unnecessary. Do we really reason about each ***it*** in this manner or do we use heuristics and shortcuts that generally work. One argument might be that we do use these mechanisms as well as heuristics. We might do this by using heuristics that allow us to quickly activate and select reasoners and inference mechanisms (i.e., the interesting parts of the playground) via probabilistic inductive inference and then subsequently perform logical inference. So, unlike prior statistical approaches that learn relationships between noun phrases and pronouns, I suggest that what is learned are relationships between imperatives and the underlying reasoning mechanisms, and that is what is activated during comprehension.

Although we propose three specific reasoning modes, we expect that there are sub-classes of situated anaphor resolution problems that cannot be resolved with just the three reasoning modes proposed. For example, if the imperatives in the discourse represent not the intent of the speaker but of another in situations where the speaker may simply be conveying a message or an order from, for example, their superior or boss. In such cases, the agent would need the capability to reason about intent of another beyond the speaker. Our approach is by no means limited to just

these three reasoning modes, and it is subject of future work to explore when and how these and other reasoning modes are triggered.

In this chapter, we did not address how these microtheory templates (partial microtheories) are learned or how the agent acquires them. The question of learning is an important one and there has been extensive research efforts in acquiring and encoding knowledge from the WWW. However, situated anaphors present a unique challenge in that much of the knowledge needed to resolve them might not be explicitly available in a dataset. Instead, this knowledge may be quite implicit acquired by the agent throughout its lifetime. We are currently exploring how an embodied agent might glean these implicit rules from experience.

We have presented the first steps towards resolving ambiguous references by reasoning with situated information available to an agent when embodied in an environment. One follow-on step for this research effort is to integrate these capabilities into a cognitive robotic architecture and attempt to empirically evaluate the system and the knowledge represented therein in real human-robot interaction scenarios. One advantage of the proposed microtheories is the use of identifiers for object constants that allow for the integration of multi-modal perceptual information about the same entity to be aggregated and reasoned with and allows for the symbols to be grounded in the robot's sensory-motor system.

2.11 Conclusion

Artificial agents interacting with humans will need to be able to disambiguate anaphoric expressions, which are used freely and frequently in discourse. To do so, we argue that the agent must consider what it *can*, *should* and be *expected* to do in a situation. In this chapter, we propose a knowledge representation scheme to formalize domain-independent and situation-specific knowledge for each of these three considerations, and a resolution strategy for using this knowledge to disambiguate object pronouns in simple imperative sentences that are situated in real-world embodied discourse. This work advances the state of the art in anaphora resolution by reframing the

disambiguation problem from being only about mention-pairs, to also being about the viability of the actions being considered as the world state evolves.

	(5) "Pass it to me." [washing dishes / cooking] $\Theta = \{knife, tomato\}$	**(4) "Put it in the bowl." [Bowl contains food]** $\Theta = \{knife, tomato, bowl\}$	**(3) "Put it down."** $\Theta = \{knife, tomato\}$
Plausibility ($\mathcal{E}_p$)	10 Stable models with "pass" $m(\{knife\}) = 0.4$ $m(\{tomato\}) = 0.3$ $m(\Theta) = 0.3$	29 Stable models with "put in" $m(\{knife\}) = 0.52$ $m(\{tomato\}) = 0.21$ $m(\{knife, tomato\}) = 0.27$	4 Stable models with "put down" $m(\{knife\}) = 0.25$ $m(\Theta) = 0.75$
Normative ($\mathcal{E}_n$)	Washing Dishes: 1 Stable Model, $m(\{knife\}) = 1.0$ | Cooking: 1 Stable Model, $m(\{tomato\}) = 1.0$	1 Stable Model $m(\{tomato\}) = 1.0$	1 Stable Model $m(\{knife\}) = 1.0$
Speaker-Intent ($\mathcal{E}_i$)	Washing Dishes: 1 Stable Model, $m(\{knife\}) = 1.0$ | Cooking: 1 Stable Model, $m(\{tomato\}) = 1.0$	1 Stable Model $m(\{tomato\}) = 1.0$	2 Stable Models $m(\{knife\}) = 0.5$ $m(\{tomato\}) = 0.5$
Combined Scores	Washing Dishes: $knife : [1.0, 1.0]$ | Cooking: $tomato : [1.0, 1.0]$	$tomato : [1.0, 1.0]$	$knife : [1.0, 1.0]$

Table 2.1: Computed uncertainties for each of the object candidates in three different scenarios. The bottom row contains the final uncertainties for the object candidates. Thus, for example, the agent is certain that the object pronoun must resolve to "tomato" when a speaker has the agent to "pass it" and the speaker was in the middle of a cooking task. Note here, for brevity we use the name of the object identifier, e.g., "knife" instead of "obj1".

Chapter 3

Interpreting Indirect Speech Acts

In this chapter, I build on the lessons learned from the previous chapter and study how a similar sense-making architecture can be employed at the sentence-level. Here, I consider a class of non-literal speech, termed Indirect Speech Acts (ISAs) that can have an ambiguous intended meaning.

3.1 Introduction

Understanding natural language utterances inherently involves resolving ambiguity associated with the meaning of linguistic expressions. Resolving linguistic ambiguity involves choosing one amongst many possible meanings, at the word level, at the sentence level, and at the discourse level. Word and sentence embeddings are current topics of active research and focus on selecting suitable word or sentence representations that best capture meaning. Current transformer architectures such as BERT [DCLT18] are used to pretrain language representations so word and sentence embeddings account for nearby linguistic context. These systems are based on two key insights that we explore in this chapter. The first insight is that word meanings (in the case of BERT) and span meanings (in the case of SpanBERT [JCL+20]) cannot be understood separate from their context, which in these cases is primarily linguistic context. Using these contextualized language embeddings has led to state-of-the-art performance in several downstream tasks. However, there is still the open

problem of how to incorporate extra-linguistic context. Specifically, it is unclear how to categorize and use knowledge derived from common sense knowledge about how the world works as well as situational knowledge about aspects of the physical environment in a systematic way for NLU.

The second insight is that the way context influences meaning can be entirely captured statistically from data. That is, a word encountered together with context in a particular way is likely to mean the same as the same word encountered in a similar linguistic context. The assumption is that systems with sufficient training data, or at least with powerful pretrained models are likely to interpret language correctly. However, Cohen [Coh19] and others have argued that pattern recognition by itself might be insufficient, and such an association might require logical reasoning beyond learning statistical similarity.

Recent work by Sarathy et al. [SS19a], explored both these insights and argued that, for the task of coreference resolution (at least with respect to imperative discourse), to interpret a pronoun an NLU system must reason about actions and change, normative obligations and social aspects, and the intent of speakers. Knowledge from situational context combined with background domain-general knowledge is needed to be reasoned with. The key idea here was that meaning is derived from choosing the interpretation that makes the "most sense" (not just in terms of statistical regularity, but in terms of coherence more broadly) given a variety of contextual factors.

We are inspired by this key idea to explore how context and reasoning can assist not only at the word level (as Sarathy et al. have done), but at the sentence level. Specifically, we address the problem of how a listener can infer the intended meaning of an utterance given contextual knowledge. Sometimes, as in **Indirect Speech Acts** (ISAs), the literal meaning of a sentence (derived from the surface form) can be different from its intended meaning [Sea75, Aus75, GL75]. For example, the utterance "can you open the door?" has a *literal meaning* of an elicitation of information from the listener regarding the listener's ability to open the door – this is because the utterance has the surface form of a question. And in fact, the speaker

may in some cases be interested in whether the hearer is capable of opening the door: if the hearer is physically disabled and the speaker is a caregiver, for example. However, very frequently, the *intended meaning* of the utterance is a request that the listener perform the action of opening the door, in which case the utterance is an ISA. As this example suggests, the proper interpretation is guided by context.

Research has shown that ISA use is common in human-human communication [Lev01], as well as in human-robot task-based dialogues [WTNS18][1]. The problem, however, is that ISAs are notoriously difficult to computationalize when their interpretation is influenced by contextual factors. Unlike past approaches to ISA interpretation, we incorporate a broader notion of context and extended reasoning capabilities to reason about the speaker's intent and beliefs.

Contributions: We provide a computational formalization of the (1) preconditions, related to the speaker's intentions and beliefs, that need to be satisfied for determining if a surface form carries indirect meaning, (2) reasoning needed to infer speaker beliefs based on domain general principles about the world (e.g., inertia), expectations of linguistic alignment, and mutual knowledge and (3) representations needed to assimilate factual contextual evidence available to the listener into the reasoning process. To make this work more tractable, we limited our focus to ISAs that have an intended meaning of a request, while their surface forms could be a question or a command. In the tradition of past work in computationally-aided linguistic analysis, these formalizations are derived using examples from several task-oriented dialogue corpora.

3.2 Approach

Our approach is based on two traditions. First, the tradition from computational linguistics of building formal representations of linguistic phenomena from corpora, guided by linguistic and cognitive scientific theories (See papers by Jiménez-Zafra et al. [JZMMUL18] and Ilievski et al. [IVS18]). We emphasize that the model that

[1]People not only use ISAs but repair ineffective ISAs with other ISAs, supporting the idea that humans have a strong preference for ISAs.

we build from this corpus analysis is a cognitive scientific model, which builds on theories from cognitive science and psycholinguistics as described in Section 3.3.

Second, the tradition of building explicit knowledge representation formalisms along with reasoners capable of deriving meaning of natural language expressions from their logical forms. Recent work in this tradition has been performed by Sarathy et al. [SS19a], Sharma et al. [SVAB15a], and Wilske and Kruijff [WK06].

Combining these traditions, we perform a corpus analysis to extract archetypal utterances from a variety of corpora; this process includes a consensus annotation of those utterance's direct and indirect meanings, as well as of the context that influences that interpretation, as described in Section 3.2.1. Then we build a formal computational model of the representations and reasoning required to perform those interpretations, ensuring that the computational model performs a complete coverage of the archetypal utterances, as described in Section 3.2.2. The resulting model is detailed in Section 3.3.

3.2.1 Corpus Analysis

We are motivated to perform a corpus analysis because of limitations in existing corpora, which include dialogue act classification and intent recognition datasets, e.g.: [LSS+17, SDB+04, JBC+98]. These corpora cannot be used as-is for reasons shown by Roque et al. [RTSS20]: because ISAs are context-sensitive, **ISA Schemas** need to be developed. These are data structures that are made up of an utterance and two different context representations; in one of the contexts the utterance is an ISA, and in the other context the utterance is not an ISA. However, these ISA Schemas do not seem to be amenable to economy-of-scale authoring techniques such as crowdsourcing, so some amount of expert authoring seems to be required. For that reason, we used a corpus analysis approach suggested by Roque et al. [RTSS20]: using real-world corpora to manually author ISAs. We proceeded as follows.

First, we obtained three existing corpora, selected to provide variation (speech vs text, virtual-world vs real-world, human-human vs human-robot, navigation vs construction) while retaining a focus on task-based communication: (1) SCARE cor-

pus [SSBFL08] in which one person helped another person navigate through a virtual environment, (2) Diorama corpus [BWTS17] in which one person directed a teleoperated robot in arranging real-world items, and (3) Minecraft corpus [NCJH19], in which one person directed another person in collaboratively building structures in a virtual environment.

Second, we manually searched the corpora (15,000+ utterances in 50+ hours of recordings) for pairs of human utterances with identical or very similar surface form, and in which one element of the pair was an ISA and the other pair was not. It was not necessary for the utterance to have been made by the same person, or even in the same corpus, because any system capable of interpreting ISAs should be domain-general. We found 20 such pairs in the corpus of 15,000 utterances, which is consistent with Roque et al. [RTSS20]'s experience of obtaining a very small yield of ISA Schemas when developing them using crowdsourced methods.

Third, having identified such pairs, we manually extracted the context details: scene, dialogue history, roles, constraints, and whether the intended meaning was indirect or direct. We identified the values of these parameters through consensus annotation. The utterance, and the context parameter/value slots, are what make up an ISA Schema. We consider these ISA Schema to be *archetypal* in the sense that they represent the set of ISAs in the corpus, including the ones that are unpaired. Our focus is on the paired ISAs, to investigate the ways that varying context varies interpretation for a given utterance.

3.2.2 Reasoner Development

Having thus identified archetypal ISA Schemas, we next developed a computational model of the logical reasoning required for their interpretation. The model is described in detail in Section 3.3, and was developed through the iterative development of logic programs, ensuring that the resulting programs achieved a complete coverage of the 20 extracted ISA Schemas. An important finding was the extent to which the knowledge, representations and reasoning requirements were domain-general and not tied to the particular corpora themselves. Thus, the reasoner we built, encod-

ing these domain-general aspects, serves as a necessary starting point for others extending this work to cover other corpora or other languages.

We encoded the representations in Answer Set Prolog, a declarative logic programming paradigm that has been used for representing knowledge and reasoning non-monotonically [GL90a]. Answer Set Prolog or Answer Set Programming (ASP) programs bear a superficial resemblance to Prolog or to production rule programs, but are different in how they work. In the ASP paradigm, logic programs are written to represent knowledge about the world. Variables in the program rules are then logically *grounded* in the constants representing a particular situation of interest. A *solver* is then used to compute answer sets in which the world definitions are logically satisfied [GK14]. We used Clingo, an ASP solver [GKK+11], for implementing the model described in Section 3.3.

ASP has several characteristics that make it useful for representing knowledge in problems where context is important [Bar03b], such as with the ISA interpretation problems we are addressing. First, ASP allows *non-monotonic reasoning*, or adding knowledge that changes currently-existing beliefs. Second, ASP represents two types of negation: *classical negation* is used to indicate propositions that are false, and *negation-as-failure* is used to indicate propositions that are considered false because they are not currently known to be true. Third, ASP allows for *choice rules* and *cardinality constraints*, which allow for the explicit encoding of world-related constraints to solutions.

One of the purposes of developing this model is to formalize the requirements and products of the reasoning involved. As described in Section 3.3, the model requires contextual evidence as input, and produces a set of candidate intents. The specifics of how this would be integrated into an embodied and situated intelligent system are beyond the scope of this chapter, but while building the model we ensured that these tasks could in principle be achievable in agents built with contemporary cognitive architectures, e.g. [RTO19, SWK+19, Lai12].

3.2.3 Related Work

Current state-of-the-art data-driven approaches to dialogue act classification employ general ML techniques like SVMs, Bayes Nets and CRFs. More recently, deep learning approaches [LHTL17, LLC+19] have shown promising results on popular datasets like the Switchboard Corpus. These approaches model dependencies between neighboring linguistic expressions using topic informtion as additional context. An advantage of these approaches is their generality across a wide range of discourses. However, a limitation in these approaches is that because their datasets possess a limited conception of context (i.e., linguistic, or at best, the topics of conversation), they do not learn how situation-specific aspects can switch the interpretation of an utterance. These systems essentially learn *conventionalized* ISAs (i.e., those whose indirect interpretation is always dominant regardless of context). We anticipate that these approaches can be combined with our approach to help improve the performance for conventionalized ISA, when extensive reasoning may not be needed.

Rule-based approaches use context-sensitive rules to logically derive indirect meanings from surface forms [BWS17, WBOS15b]. These approaches address some of the limitations of the purely data-driven techniques by explicitly accounting for context. However, they generally do not reason over multiple aspects of context. Moreover, a single chunked rule (suggested by these approaches) is not domain-general and does not provide much in way of explanation for the intent recognition process. Asher and Lascarides [AL01] provided a formal account of ISAs but were primarily focused on the semantics of conventionalized ISAs. Our approach addresses these limitations and is significantly more elaboration-tolerant in allowing for small factual variations to switch the interpretation.

Plan-based approaches [PA80, HA89, GC99] reason about ISAs using a set of plan recognition techniques to infer the goals of a speaker. None of these plan-based approaches systematically enumerate the different types of context, as our approach does. These approaches do not incorporate the influence of the speaker beliefs about plausibility, preference and normativity, which we have argued to be important and

reflective of assumptions about the listener's capability and preferences. That said, these approaches could be integrated with our approach as one way to infer speaker intent.

The closest line of work is that of Wilske and Kruijff [WK06], who proposed an ISA interpretation architecture for human-robot task-based interactions. They discuss the notions of "feasibility," "mode of operation," and the use of interactional history, which appear to parallel our notions of capability, contextual role, and dialogue history, respectively. However, like the other existing approaches, they do not consider how a listener models the speaker's beliefs. For example, in their work, if a listener considers an action to be feasible, it will take it as a request and immediately perform the action, independent of whether or not the speaker believed the listener to be capable.

3.3 Reasoning for ISA Interpretation

Brown [Bro80] provided an early descriptive account of ISAs, in which she elaborated on their form, what is required to choose one form over another, and what is needed for these surface forms to carry indirect meaning. We propose that these notions are useful not just descriptively to study ISAs as inherently interesting phenomena, but also computationally as a rubric to aid in understanding and interpreting them. We computationalize the intent preconditions, and incorporate Brown's descriptive analysis of the surface form of ISAs into the architecture by allowing the surface form to dictate the necessary burden of proof required by the reasoners. Our approach, however, extends Brown's work significantly, by proposing bridge rules, reasoner-specific rules, and underlying definitions for the reasoners and their functions. Here, we systematically operationalize the principles outlined by Brown, first by formalizing the preconditions as rules in the reasoners, and then by using the surface form to influence which rules are selected for a particular utterance.

Figure 3.1, which shows an example of the reasoning required for interpretation, serves as an overview of our approach. The Preconditions of an interpretation

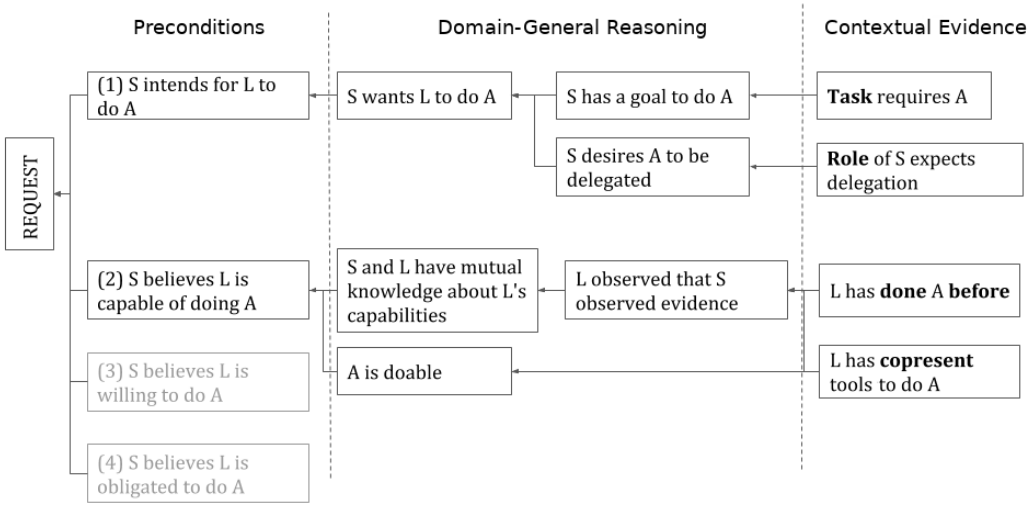

Figure 3.1: Types of Reasoning required to interpret the utterance "Can you do (action) A?" made from a Speaker S to a Listener L. To select the indirect meaning of a REQUEST for physical action (instead of the direct meaning of an ASK for information), L must satisfy four **Preconditions**. However, L does not have direct access to S's intentions and beliefs, and must perform **Domain-General Reasoning** (using rules about actions, mutual knowledge, social norms and obligations, and speaker preferences) based on **Contextual Evidence** (involving task, role, interactional history, and copresence). Reasoning and Evidence for Preconditions 3 and 4 follow the same pattern as for Precondition 2.

are described in Section 3.3.1, the Domain General Reasoning is described in Section 3.3.2, and the Contextual Evidence required for that reasoning is described in Section 3.3.3.

3.3.1 Preconditions for Interpretation

As shown in Figure 3.1, each underlying communicative intent has a set of preconditions that must be met in order for a speaker to produce a surface form that conveys this intent. The theoretical justification for this comes from Brown's descriptive analysis of ISAs, which extends work by Gordon and Lakoff [GL75]

As shown in Figure 3.1, a *Request* for a physical action (e.g., "Can you do action A?") intent must satisfy the following preconditions: (1) that the speaker S intends that the listener L **do the action** A, (2) that S believes that L plausibly **has the capability** to do A, (3) that S believes that L **is willing to** do A, (4) that S believes that L is normatively **obligated** to do A. More generally, these four

preconditions also apply to other intended meanings such as *Ask* for information, where A is a speech act rather than a physical action.

Brown also suggested that the specific surface form of the ISA is influenced by the *better-knowledge principle*, which states that the direction of information flow is guided by who, S or L, has better knowledge about a particular proposition in the ISA.[2] So S may select a "can you do A?" question when S needs more evidence for satisfying precondition 2, and a "will you do A?" question when S needs more evidence for satisfying precondition 3.

3.3.2 Domain-General Reasoning

As shown in Figure 3.1, several sets of axioms and commonsense rules are used to test whether preconditions hold in a particular interaction between a speaker and a listener.

First we consider a set of **Domain-General Axioms** that allow an agent, regardless of situation, to reason about action and change, mutual knowledge, and linguistic alignment.

One subset of axioms involves *Reasoning about Actions and Change*. A basic component of the logical evaluation of ISAs is being able to track and infer things from the ongoing changing context. The Event Calculus (EC) is especially suited to reasoning about events and provides some axioms that are domain-independent to help guide this discussion [KS89].

Formula	Meaning
$holds(F,T)$	Fluent F holds at time T
$happens(A,T)$	Ontic Action A occurs at time T
$initiated(F,T)$	Fluent F is initiated at time T
$terminated(F,T)$	Fluent F is terminated at time T

To maintain fluents once they are initiated and unless they are terminated, we have the following two rules:

```
holds(F,T+1) :- initiated(F,T),time(T).
```

[2] See also [DR12] for additional discussion on "knowledge gradients".

```
holds(F,T+1) :- holds(F,T), not terminated(F,T), time(T).
```

Another subset of axioms involves *Reasoning about Mutual Knowledge*. One challenge in interpreting ISAs is understanding not only what the speaker knows, wants and believes, but also what the speaker thinks the agent (listener) knows, and what the speaker thinks the listener thinks the speaker thinks, and so on, ad infinitum. Herbert Clark described a heuristic-based mutual knowledge induction schema to terminate this otherwise infinite line of reasoning [CM81]. Under this schema if the agent knows that a grounds $F1$ holds, and that these grounds support a proposition $F2$, and that they know that the speaker P believes the grounds $F1$ to hold, then the agent has mutual knowledge with the speaker P. The agent can establish that if it knows that the speaker P observed the grounds (any proposition F), then they are likely to believe it holds. Moreover, if the grounds support a proposition, then, at least from the agent's perspective the supported proposition $F2$ holds as well.

Formula	**Meaning**
$observed(P, F, T)$	P was seen as observing F at time T
$bel(P, F)$	P believes fluent F holds
$supports(F1, F2)$	Fluent $F1$ can provide grounds for fluent $F2$ to hold
$mutual(P, F2)$	There is mutual knowledge with P about fluent F

```
holds(mutual(P,F2),T) :- holds(supports(F1,F2),T), holds(F1,T),
    holds(bel(P,F1),T), time(T).
holds(F2,T) :- holds(F1,T),holds(supports(F1,F2),T), time(T).
holds(bel(P,F),T) :- observed(P,F,T), time(T).
```

Another subset of axioms involves *Reasoning about Linguistic Alignment*. Under the theory of alignment, there is a mapping from an utterance's surface form (which could be a question, statement, command, etc.) to a direct meaning (ask,

assert, request, etc.) [AL01]. Conventionally, questions map to *Asks*, statements map to *Asserts* and commands map to *Requests*. Surface forms can also have indirect meanings and thus we need to be able to distinguish these, which we do via the following alignment axioms:

```
alignment(question,ask).
alignment(command,request).
alignment(statement,assert).
meaning(M) :- alignment(_,M).
surfaceForm(S) :- alignment(S,_).
direct(U,M) :- hasSurfaceForm(U,S), alignment(S,M), meaning(M), surfaceForm(S),
    utterance(U).
indirect(U,request) :- not direct(U,request), utterance(U).
```

Finally, a set of axioms we assume but do not explicitly discuss are *Uniqueness-of-Names* axioms which ensure that predicate relations do not clash.

Next, we consider the **Burdens of Proof** required for domain-general reasoning. As a preliminary, consider the utterance "can you do action A?" There are two types of action underlying this utterance: *ontic action* specified as A and referring to the physical act, and *epistemic action*, which in this case where the utterance is a question, is to retrieve and respond with information about A.

Formula	**Meaning**
$onticAction(U, A)$	Utterance U contains an ontic action A
$epistemicAction(U, M, A, Q)$	Utterance U contains an epistemic action A that holds when U has meaning M, and that pertains to proposition Q
$uttered(P, U, T)$	Interlocutor P uttered U at time T
$goal(P, A)$	Interlocutor P has the goal to do action A

As another preliminary, consider four different lines of reasoning relevant to ISA interpretation – speaker intent, plausibility, preference, normativity – each of which attempts to connect a set of facts to whether or not a particular intent holds.

Each line of reasoning can be thought of as asking different questions of the facts of a given situation, as described below. These four lines of reasoning correspond directly to Brown's four preconditions for an intent to hold. But since we know from the theory of alignment that certain defaults hold we can adjust our burden of proof to make it easier to satisfy the default (or direct) interpretation than it would to satisfy the indirect interpretation. In other words, for a question "Can you install linux on your computer?" we can state that the burden to satisfy the interpretation of this utterance being an *Ask* for information should be lower than it being a *Request* for action because the surface form of it is a question and the direct meaning of questions is typically that of an *Ask* for information. One way to specify this default burden is through the use of both classical negation and negation as failure in nonmonotonic reasoning. We explain these burdens in more detail below.

First, the *Speaker Intent Burden of Proof* asks the question: does the speaker *want* the listener to have the goal of doing action (ontic or epistemic) A? If there is *no* proof that the speaker did *not* want the listener to do the action A, where A is an epistemic action, then the speaker intended for the listener to interpret the direct meaning M. For example, in the case of the utterance "Can you do action A?" if there is no proof that the speaker did not want the listener to answer the question about A, then the speaker likely intends for this to be a question and expects an answer.

```
holds(intends(P1,U,M),T) :- not -wants(P1,goal(P2,A),T), uttered(P1,U,T),
    epistemicAction(U,M,A,_), direct(U,M), holds(speaker(P1),T),
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T),
    indirect(U,M2).
```

If there is positive evidence that the speaker wanted the listener to do the ontic action A, the speaker intended for the listener to interpret the indirect meaning M. For the above example, if there is positive proof that the speaker wanted the listener to perform the physical action A, then the speaker intends for the action to

be done.

```
holds(intends(P1,U,M),T) :- wants(P1,goal(P2,A),T), uttered(P1,U,T),
    onticAction(U,A), indirect(U,M), meaning(M), holds(speaker(P1),T),
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T),
    direct(U,M2).
```

In addition, we also provide some ancillary rules to help establish that once an action is performed, the intent that provoked this action is terminated. Moreover, once the intent is formed, the action to execute this intent is performed.

```
terminated(intends(P,U,M), T) :- happened(A,T),
    holds(bel(P,responsive(A,U,M)),T).
happened(A,T+1) :- holds(intends(P1,U,M),T), uttered(P1,U,T), action(U,A).
```

Second, the *Plausibility Burden of Proof* asks the question: does the speaker believe that the listener is *capable* of doing action A? Burden of proof rules similar to that of Speaker Intent can be formulated for these other reasoning modes as well. For example, for Plausibility reasoning, the key question is whether there is mutual knowledge about the capability of the listener.

```
holds(intends(P1,U,M),T) :- not -holds(mutual(P1,capable(P2,A)),T),
    uttered(P1,U,T), epistemicAction(U,M,A,_), direct(U,M),
    holds(speaker(P1),T), holds(listener(P2),T), time(T), not
    holds(intends(P1,U,M2),T), indirect(U,M2).
holds(intends(P1,U,M),T) :- holds(mutual(P1,capable(P2,A)),T), uttered(P1,U,T),
    onticAction(U,A), indirect(U,M), meaning(M), holds(speaker(P1),T),
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T),
    direct(U,M2).
```

Third, the *Preference Burden of Proof* asks the question: does the speaker believe that the listener is *willing* to do action A? This has rules analogous to the Plausibility Burden of Proof, with $willing(P2, A)$ replacing $capable(P2, A)$.

Finally, the *Normativity Burden of Proof* asks the question: does the speaker believe that the listener is *normatively obliged* to do action A? This has rules analogous to the Plausibility Burden of Proof, with $obligated(P2, A)$ replacing $capable(P2, A)$.

3.3.3 Contextual Evidence

As shown in Figure 3.1, facts and perceivable situation-specific aspects of a context connect to the burdens of proof. To represent this, we can specify some optional ancillary rules as well as commonsense knowledge that is associated with each dimension of fact. First, the **task**-related constraint aspect of context, which relates to limitations imposed by the nature of the action or of the task, as understood by the interactants. Second, the interactant **roles** aspect of context, which relates to the duties, expectations, and obligations of the interactants derived from their respective social or occupational status; whether one interactant has authority over the other, for example. Third, the **interactional history** aspect of context, which relates to the utterances and actions that have been performed in recent memory by the current interactants. Finally, the **co-presence** aspect of context, which relates to the objects and participants that are present when the utterance is made. The first three of these are motivated by dialogue context models, as for example surveyed by Jokinen and McTear [JM09]. The last of these is motivated by Clark's copresence heuristics [CM81].

The contextual evidence is used by domain-general reasoning in the following way. To determine if a speaker wants (or does not want) the listener to do an action A as analyzed by the speaker intent reasoner, we will need to consider whether the speaker intends to have the action accomplished in the first place (i.e., it is the speaker's goal to get the action done) and whether the speaker would wish to delegate this action to the listener. We can establish this line of reasoning with the following three rules. The first rule states what is needed for the speaker to want

the listener to do an action. The second and third rules state how this "want" can be negated.

```
wants(P1,goal(P2,A),T) :- holds(goal(P1,A),T), holds(delegate(P1,P2,A),T),
    holds(mutual(P1,isAvailable(P2)),T), action(_,A).
-wants(P1,goal(P2,A),T) :- -holds(goal(P1,A),T),
    holds(mutual(P1,isAvailable(P2)),T), action(_,A).
-wants(P1,goal(P2,A),T) :- -holds(delegate(P1,P2,A),T),
    holds(mutual(P1,isAvailable(P2)),T), action(_,A).
```

The following commonsense default rules specify how contextual evidence contributes to reasoning:

```
% TASK-RELATED
holds(goal(P,A),T) :- holds(requires(K,A),T), holds(currentTask(K),T),
    holds(speaker(P),T), onticAction(_,A).
-holds(goal(P,A),T) :- holds(currentTask(K),T), holds(requires(K,Q),T),
    epistemicAction(_,_,A,Q), holds(speaker(P),T).

% ROLE EXPECTATIONS
holds(delegate(P1,P2,A),T) :- holds(role(P1,leader),T),
    holds(role(P2,follower),T), onticAction(_,A).
-holds(delegate(P1,P2,A),T) :- holds(role(P1,follower),T),
    holds(isPresent(P2),T), onticAction(_,A).
:- holds(role(P,follower),T), holds(role(P,leader),T).

% INTERACTIONAL HISTORY
holds(delegate(P1,P2,A),T) :- happened(A,S), time(S), S<T,
    holds(speaker(P1),T), holds(listener(P2),T), onticAction(_,A).

% CO-PRESENCE
observed(P1,capable(P2,A),T) :- observed(P1,responsive(A,_,_),T),
    holds(isPresent(P2),T), onticAction(_,A).
```

```
holds(delegate(P1,P2,A),T) :- holds(mutual(P1,capable(P2,A)),T),
    onticAction(_,A).
```

The above discussion specified the evidence and reasoning with respect to Precondition 1 (speaker intent). The Preconditions 2-4 involve analogous reasoning, which is omitted for reasons of space.

Finally, the following example shows how speaker observations update the interactional history evidence over several dialogue turns. Of particular interest is the utterance at time step 7, which has a literal meaning of an *Ask* for information but an intended meaning of a *Request* for action, making it an ISA.

```
% Narrative
% t=1. p1 says: "You're gonna go forward"
holds(intends(p1,u1,request),1).
% t=2. p1 says: "stop"
happened(goForward,2).
observed(p1,responsive(goForward,u1, request),2).
holds(intends(p1,u2,request),2).
% t=3. p1 says: You're gonna turn right 90 degrees
happened(stop,3).
observed(p1,responsive(stop,u2,request),3).
holds(intends(p1,u3,request),3).
% t=4. p1 says: Can you grab the second box
happened(turn,4).
observed(p1,responsive(turn,u3,request),4).
holds(intends(p1,u4,request),4).
% t=5. p1 says: Can you push forward
happened(grabItem,5).
observed(p1,responsive(grabItem,u4, request),5).
holds(intends(p1,u5,request),5).
% t=6. p1 says: just go forward
happened(pushItem,6).
```

```
observed(p1,responsive(pushItem,u5, request),6).
holds(intends(p1,u6,request),6).
% t=7. p1 says: Can you grab the box?
happened(goForward,7).
observed(p1,responsive(goForward,u6, request),7).
uttered(p1,u7,7).
hasSurfaceForm(u7,question).
onticAction(u7,grabItem).
```

3.4 Generating ISA Schemas - Building an ISA Corpus

In the previous sections, we have argued that intelligent systems that interact with people in real-world environments need to be able to use context to determine whether an utterance should be interpreted literally, or as part of an ISA. An important part of developing such a capability is testing it. But doing so requires answering several difficult questions regarding how to collect and represent such content. In this section, we focus on the task of corpus building and describe one way in which useful corpora can be constructed.

Specifically, we describe an approach to testing ISAs that is derived from relevant work in using collections of test problems to track progress in systems that perform reasoning tasks. We present a formal representation of ISA Schemas required for such testing, including a measure of the difficulty of a particular schema[3]. We develop an approach to authoring these schemas using a combination of corpus analysis and crowdsourcing, to maximize realism and minimize the amount of expert authoring needed. Finally, we describe several characteristics of collected data, and potential future work.

[3]The plural of *schema* is *schemas* [OED19]; we adapt this term from *Winograd Schemas* as described in Section 3.5

3.5 Some Related Work in Corpus Building for Knowledge-Based Agents

AI researchers have developed a variety of language-processing tests that require reasoning about knowledge [SGC19]. Examples include COPA [RBG11] and RTE [RMMG08]. Trichelair et al [TETC19] describe some problems associated with the datasets used for such tests, including limited size and the predictable structure of their examples. The terms **corpus** and **corpora** are used for these datasets by e.g. Levesque et al. [LDM12a] and Morgenstern et al. [MDO16], even though this usage differs somewhat from the traditional usage in linguistics, i.e. referring to a collection of texts and conversations.

One approach we find particularly inspiring is that of *Winograd Schemas* (WS), which are used to test a system's ability to perform anaphora resolution. The following is an example, reformatted from Levesque et al. [LDM12a], of a WS problem:

> Statement: The trophy doesn't fit in the suitcase because it's too [big / small].
>
> Question: what is too [big / small]?
>
> Answer: [the trophy / the suitcase].

This example shows how schema are actually made up of two halves, called *options*. In Option 1, "big" is selected for the statement and question and the correct answer is "the trophy;" in Option 2, "small" is selected for the statement and question and the correct answer is "the suitcase." When a schema is presented, typically only one option is shown, and the system or person being tested needs to select the right answer.

We extract several important lessons from the experience of Winograd Schemas. First, regarding **levels of difficulty**: WS researchers distinguish between *easy* and *hard* WS problems. Easy WS problems are those that can be solved by (1) *statistical*

correlations e.g. simply observing whether the query words co-occur more frequently with one of the possible solutions, (2) *selectional restrictions* in which the answer can be determined just by using definitions of the options, and (3) other simple *syntactic cues* [Ben15]. Hard WS problems require reasoning about knowledge. In the example above, determining what the pronoun "it" refers to requires reasoning that in general if object A does not fit in object B, then object A is bigger than object B. The words "big" and "small" are equally applicable to "trophy" and "suitcase," so a system that only uses statistical correlations will do no better than chance.

This leads to the second important lesson we extract from the experience of Winograd Schemas, regarding **the importance of alternate options.** Having each schema be made up of two options enables the hard problems that cannot be solved by statistical correlations, because the statements in each option only vary slightly. (WS are often authored such that they only vary by a single word, though in some cases this may be several words.) This also ensures that the corpus of schemas is testing for those slight variations that create large changes in meaning.

This contributes to the third important lesson we extract from the experience of Winograd Schemas, regarding **the difficulty of developing a corpus of such schemas.**

The ideal solution would be to find several such paired halves in naturally-occurring data, but to the best of the community's knowledge there is no such source of naturally-occurring data. One obvious solution is to have experts construct them, but this is time-consuming and potentially leads to unrealistic data. These challenges are spelled out in more detail in Section 3.7 Our general approach to solving these problems is described in Section 3.8 An example corpus development is given in Section 3.9 But first, the next section provides our formal definition of ISA Schemas.

3.6 The "ISA Schema"

Imagine a person injures their leg and goes to a doctor. In the doctor's office, the doctor says that they will begin by asking about the extent of the injury. The doctor

says: "Can you run?" In this context, the utterance is clearly a question asking about the patient's capability.

After a few more questions, the doctor determines that the patient can in fact run, and that it is safe to test the extent of the patient's injury. The patient is taken to a treadmill in the exercise room, and the doctor says: "Can you run?" In this context, the utterance is clearly a request that the patient begin running.

So the utterance is identical in both cases; only the context has changed. This contextual change can be informally represented as follows.

Utterance:	*Can you run?*
Context 1:	*a doctor talking to a patient,*
	in a doctor's office,
	to collect information for diagnosis.
Context 2:	*a doctor talking to a patient,*
	in the exercise room,
	to test physical capability.

There are several cues for interpreting the utterance. First, *role* of the speaker and of the hearer. Second, *location*: in an office it is unusual to suddenly begin running, whereas on a treadmill it is perfectly natural. Third, *task*: in Context 1 the task is to collect information about the injury, whereas in Context 2 the task is to test physical capability; it is therefore more plausible that the utterance in Context 1 is asking about an ability and in Context 2 is requesting an action. Fourth, *copresent items*, such as a treadmill. Finally, *interaction history*: in Context 2 the patient has already answered the question and the doctor has already diagnosed that it was safe for the patient to run. All of this suggests that the utterance in Context 1 is asking about an ability, and the utterance in Context 2 is requesting an action. The literal meaning of the utterance is asking about an ability, so in Context 2 the utterance is an ISA.

This example shows how, for a given utterance, variations in context produce

variations in interpretation. We therefore define an **ISA Schema** as

$$S_{ISA} = (u, l, c_{1...n}, i_{1...n})$$

where u is an utterance, l is the literal meaning of u, $c_{1...n}$ is a set of contexts in which u is made (where a context is defined by a set of feature pairs), and $i_{1...n}$ is the intended meaning of u for contexts 1 to n. So for a given context c_x, u is an ISA iff $l \neq i_x$.

Figure 3.2 shows an example ISA Schema. This schema includes context features including the task at hand and the feature role, with a different intended meaning i for each context. The utterance is not an ISA in context 1 (because its literal meaning is the same as its intended meaning) and the utterance is an ISA in context 2.

[l]utterance	*Can you run?*
literal-meaning	*ask-ability*
context-1	[l]task: *collect-information*; speaker-role: *doctor*; hearer-role: *patient*; location: *doctors-office*; copresent-item: *chair*; interaction-history: *none*
intended-meaning-1	*ask-ability*
context-2	[l]task: *test-capability*; speaker-role: *doctor*; hearer-role: *patient*; location: *exercise-room*; copresent-item: *treadmill*; interaction-history: *has-been-diagnosed*
intended-meaning-2	*request-action*

Figure 3.2: Example ISA Schema, where the intended meaning of an utterance varies based on context, and an Indirect Speech Act is performed in context 2 when the intended meaning is not equal to the utterance's literal meaning.

We are interested the difficult cases in which context affects intended meaning. So we additionally require that (1) at least one of the contexts must be an ISA (i.e.

$l \neq i$ for that context) and at least one of the contexts must *not* be an ISA (i.e. $l = i$ for that context), (2) a system to interpret ISA should look at each context of the ISA separately, e.g. first (u, l, c_1) and then (u, l, c_2). This follows from the experience of WS Schemas as described in Section 3.5

3.7 Techniques for Developing a Corpus of ISA Schemas

Having defined S_{ISA}, we next need to define how these S_{ISA} can be obtained.

From the relevant work in developing corpora for reasoning tasks, we can identify several approaches. This section will show how any technique for developing a corpus of ISA Schemas will involve actions on a continuum involving trade-offs between naturalness, accuracy, and scalability. After this section describes each of these techniques and the trade-offs they involve, the next section describes how we developed an approach that appropriately balances these trade-offs.

3.7.1 Expert Authoring

First, we ourselves could author $S_{ISA}s$; this is the approach that was initially taken by WS researchers.

The challenge is that this approach is time-consuming, although regarding WS, Levesque et al [LDM12a] argue that the amount of effort required to author problems, while not negligible, is not insurmountable either. Indeed, Levesque et al. manually authored 273 schemas, although their approach does not scale to create thousands of schemas. However, once these expert-authored schemas are available, they have been shown to be translatable into other languages [AS17b].

Another challenge is that this approach produces schemas that are potentially not representative of material that would be found "in the wild." However, Levesque et al [LDM12a] argue that WS problems can be derived from natural data, and that in practice, WS problems can be combined with more-easily-obtained *Pronoun Disambiguation* problems, which are like one half of a WS, thereby being "easy" problems by definition [LDM12a].

One of the disadvantages of having experts perform the authoring is that they may subconsciously tailor the resources authored based on the needs of their research. One of the advantages is that the resources are more likely to be high-quality, as the experts will have a strong idea of what the schema represents, and have spent time considering what they should look like. No matter what type of authoring approach is used, the corpus benefits from having the experts examine at least a subset of the corpus to ensure the appropriate level of quality is being maintained.

3.7.2 Non-Expert Authoring

A second approach we could use is to have non-experts author S_{ISA}.

This has been done with WS in several different ways. Rahman and Ng [RN12] had 30 undergraduate students create WS, resulting in 941 schemas. A more scalable approach was taken by Sakaguchi et al [SBBC19], who used the Amazon Mechanical Turk crowdsourcing platform to author a set of over 40,000 schemas.

The advantage of having non-experts author schemas is that it removes the authoring burden from researchers, thereby providing a more scalable approach that can generate larger corpora. It also lessens any potential subconscious bias. The disadvantage is that the non-expert authors are still performing a somewhat abstract language task, and therefore may produce results that are not necessarily tied closely to the reality of their daily language use. Another disadvantage is that non-experts are more likely to produce noisy output.

3.7.3 Non-Expert Validation

Regardless of the approach used to develop a schema, the corpus benefits from validation by examining the corpus. This can be done by experts as described above, but using experts does not scale to larger corpora.

In developing their WS, Bender et al [Ben15] conducted a study of how correctly humans interpreted a corpus of WS, and found that participants received an average score of 92%. That study also helped to identify ill-formed problems that

had been human-authored, but which were not evident at the authoring stage because the author had access to both options of the WS (whereas the validator only had access to one of the options.)

In the same way, Zellers et al [ZBSC18] used Amazon Mechanical Turk to have human non-experts validate a multiple-choice corpus of grounded commonsense inference statements.

The advantage of this approach is that it provides a "reality check" on the quality of the data. The disadvantage is that the non-expert validators may incorrectly process the corpus elements, either allowing substandard schemas or removing acceptable schemas.

3.7.4 Defining a Task to Collect Data

One intriguing possibility is to have subjects work on a task that "naturally" produces $S_{ISA}s$. This would require defining a task that, in the course of execution, would require someone to use a given utterance (1) in a context in which the utterance's literal meaning was intended and then (2) in a context in which the utterance's literal meaning was **not** intended.

Collecting $S_{ISA}s$ in this manner would have the advantage of being more natural than corpus-based non-expert collections. However, the disadvantage is that defining such a task is prohibitively difficult without a well-defined notion of what S_{ISA} are, and how to characterize their quality. We therefore reserve this approach for future work.

3.7.5 Extracting from Existing Corpora

This approach is a variation of the previous technique: rather than create a task to collect data, we could extract schemas from an existing corpus. However, as described in the previous technique, there is no single corpus that naturally contains the schemas that we need. We therefore looked through sets of corpora to see if an utterance in one corpus might have an identical or near-identical utterance in

another corpus, in different contexts, that we could combine together to make an S_{ISA}.

Indeed, this enabled us to perform expert authoring of several schemas that were thus tied to existing corpora. However, because this approach used expert authoring, it did not scale. We considered using automated approaches to extract the $S_{ISA}s$, but we identified several challenges: (1) automatically identifying when utterances were "similar enough" (i.e. whether the utterances had to be exactly similar word-for-word), (2) automatically extracting the context information in a way that was comparable across corpora, and (3) automatically determining the intended meaning of the utterances. Indeed, accomplishing (3) was one of the motivations for developing an ISA corpus to begin with.

Using existing corpora promises to result in highly-realistic schemas. Therefore, although it is currently infeasible to extract schemas entirely from corpora, the next section shows how we use data (such as utterances and contexts) from existing corpora whenever possible.

3.8 A General Approach to Developing a Corpus of ISA Schemas

As described above, any step in developing a corpus of ISA Schemas will involve actions on a continuum between explicit authoring and natural occurrence, with explicit authoring providing scalability problems. This section therefore defines a general approach which maximizes both scalability and ties to realistic data. The following section describes an execution of this general approach.

Given: the definition of $S_{ISA}s$ presented in Section 3.6

Step 1: we identify prospective *utterances* for the $S_{ISA}s$. Rather than author these ourselves, we automatically extract them from a corpus. We review the utterances only long enough to remove any that, for example, have unusual characters. In other words, to help ensure scalability, we do not individually examine them.

Context Feature	Suggested Context Values
Location	School, Store, Street, Restaurant, Offices, Car on the road, Pool, University, Hotel, Home, Phone call, Nondescript
Task	getting contact information, looking for a hotel, asking about a book, mailing a letter, talking about a show, general dialogue, planing a meeting, planning errands, looking for a rest stop, grocery shopping
Speaker Role	Volunteer, Administrator, Tourist, Tour Guide, Local, Customer, Employee, Driver, Assistant, Waiter, Teacher, Student, Interviewer, Interviewee, Translator, Public Servant
Hearer Role	Volunteer, Administrator, Tourist, Tour Guide, Local, Customer, Employee, Driver, Assistant, Waiter, Teacher, Student, Interviewer, Interviewee, Translator, Public Servant
Copresent Item	car, buildings, hotel, theater, letter, package, phone, computer, toy, stamps, stone, food, utensil, notepad, pictures, clothes, fabric, money
Interaction History	this is the beginning of the interaction, they have talked about a plan to accomplish the task, the person is beginning an order, they have already talked about the copresent item, they have talked about events related to the task, they have determined that the person wants a specific item, the person has just arrived at the location, they are having difficulty communicating

Figure 3.3: Suggested Context Features and Values, manually extracted from corpora by the authors

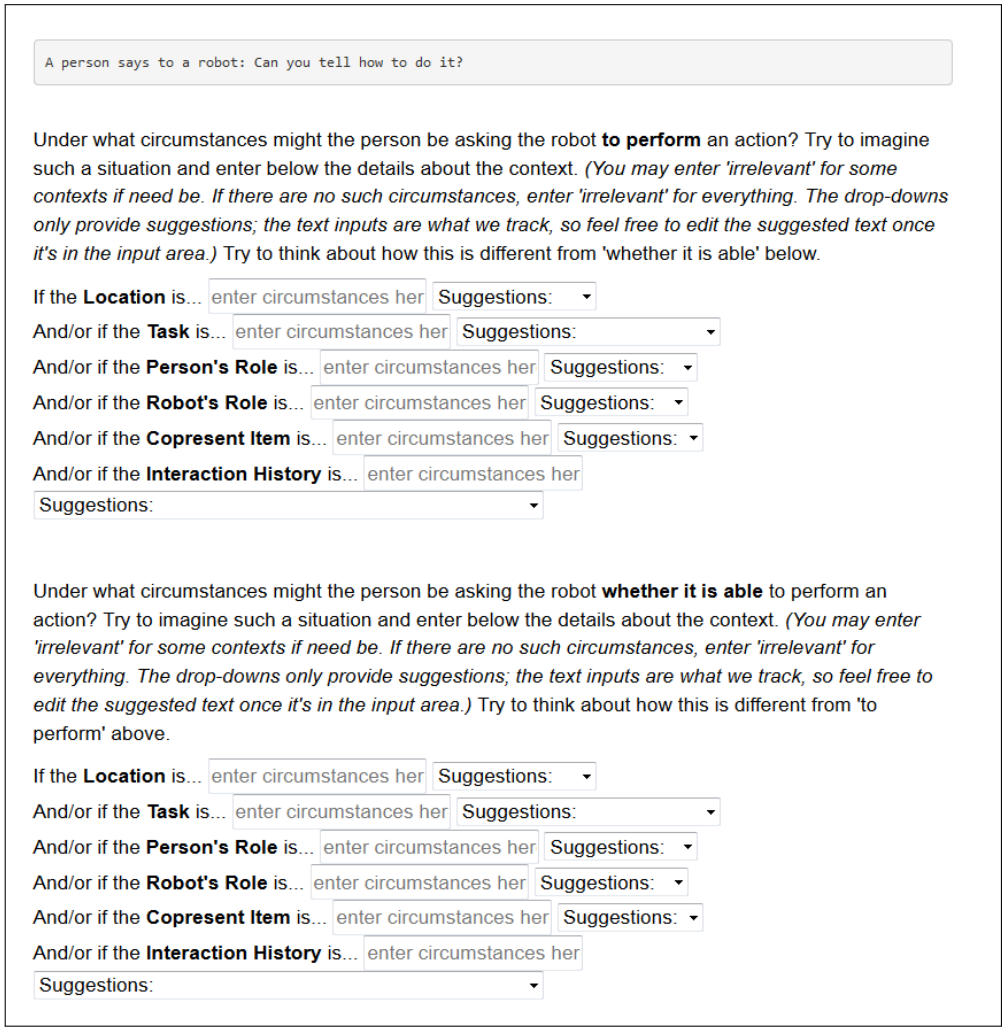

Figure 3.4: Example of the web-based GUI used for non-expert authoring of ISA Schemas.

Step 2: for each context feature, we author several *suggested context values* by extracting them from corpora. This involves expert authoring, but in terms of scalability the number of these suggested context values is constant per corpus (i.e. a corpus with 100 schemas may have the same number of suggested context values as a corpus with 10,000 schemas.)

Step 3: we use non-experts to *author* values for the context features, guided by the the suggested contexts, enabled by crowdsourcing as described in Section 3.7.2

Step 4: we use non-experts to *validate* the schemas that have been authored, enabled by crowdsourcing as described in Section 3.7.3

Step 5: we use a limited amount of expert authoring to produce the best-validated schemas. To maintain scalability, we minimize the amount of expert authoring, such as only requiring experts to act as "tie-breakers" as described in the example below.

3.9 Example Development of a Corpus of ISA Schemas

We now describe a corpus development whose goals were the following. We wanted to ensure that we had defined an approach to collecting a corpus of S_{ISA}s in a way that minimized authoring effort while maximizing realism and maintaining scalability. For our first effort, we also wanted to produce a corpus that was small enough to be closely examined, while ensuring that we used scalable approaches. From our familiarity with several corpora [LSS+17, EM17, SYC+19] we decided to focus on utterances of the form "Can you...?"

For **Step 1**, we used a script to extract all utterances from the DailyDialog corpus [LSS+17] that took the form "Can you...?", and we organized these utterances in a hash table with a key of the first three words to nominally cluster like sentences. We then randomly extracted 250 sentences and manually discarded all sentences that on a quick reading seemed to have unclear grammar or were not very understandable. This left a set of 215 utterances.

For **Step 2**, we authored suggested context features as follows. First, we

took a sample of 20 utterances from the 215 utterances identified in Step 1. Next, we identified 2 possible contexts for each of these 20 utterances, and collected all unique context features. The suggested context features we identified are shown in Figure 3.3.

For **Step 3**, we constructed a web-based GUI, shown in Figure 3.4, that used the elements of Steps 1 and 2. One of the utterances identified in Step 1 is shown at the top of the GUI. The context features from Step 2 are provided as suggestions in a drop-down menu, which when selected, populates the editable text input next to each context feature name. As shown in Figure 3.4, the authors are encouraged to modify the editable text or to enter text that was not suggested. They also may enter "irrelevant" for any given feature, or to enter "irrelevant" for all features if they cannot imagine an appropriate scenario.

We used this GUI to collect data using the Amazon Mechanical Turk and psiTurk [MMM+12] crowdsourcing tools, adhering to the oversight of an Institutional Review Board. Before being shown the GUI in Figure 3.4, the participants were given the following overview:

> On each of the following pages, you will be shown a description of a human-robot interaction. On the same page you will see several questions for you to answer related to that description.
>
> We are interested in how context can change interpretation. Imagine a doctor says to you: "Can you move your arm?" In some circumstances this might be a request for you to perform an action, such as moving your arm out of the way. In other circumstances it might be a question about whether you are able to perform an action, such as after a serious injury. That is why we are asking questions about circumstances.

We used the word "circumstances" to help make the idea of "context" intuitive to the participants. We framed the "Can you..." utterance in the context of a robotic interaction in part because that is the scenario that we are interested in, and in part to provide a concrete basis for the annotators to consider.

The GUI shown in Figure 3.4 was used to generate 92 $S_{ISA}s$ (80 schemas authored by 20 participants authoring 4 schemas each, and 12 additional schemas being authored by participants who did not complete the process of authoring a set of 4.)

As a first characterization of the collected data, we began by considering the difficulty of the schemas collected. As described earlier, WS are characterized as either "easy" or "hard" depending on how amenable they are to statistical analysis, which is affected by the limited number of words that are changed between the two options. We therefore define the **contextual difficulty** of an S_{ISA} as the number of nonzero context features that are non-identical between its two contexts. So if two contexts, with different intended meanings, are identical except for 1 feature, then it has a contextual difficulty of 1, which is the maximum difficulty.

Contextual difficulty is defined for a nonzero number of contexts because when two contexts sets are identical but have different intended meanings, this indicates that the meaning is ambiguous. For our current purposes, we are taking note of when such ambiguities occur but we leave a full exploration of ambiguities in S_{ISA} for future work. Additionally, the existence of schemas whose two halves are identical could also be the result of authors who did not understand the task.

The number of examples of schemas for each contextual difficulty level, for the first set of collected data, is shown in Table 3.1. In addition, 14 0-level (i.e. ambiguous) schemas were identified. Note that the measure of contextual difficulty is affected by the existence of the "irrelevant" keyword which the schema creators were instructed to use; when measuring contextual difficulty "irrelevant" is considered a variable which matches any other word.

For **Step 4**, we developed and used a GUI as shown in Figure 3.5 to perform non-expert validation through crowdsourcing. The purpose of this was to determine the extent to which human perception of intended meaning corresponded with the authored intended meaning.

Each schema authored in the previous step was split into two halves based on their different context. Then an annotator was presented with the half-schema in the

Contextual Difficulty	Examples Found
1	14
2	14
3	10
4	13
5	12
6	15

Table 3.1: Difficulty Level of Schemas, based on differences in context features, in a collection of 78 $S_{ISA}s$, where 1 is most difficult and 6 is least difficult.

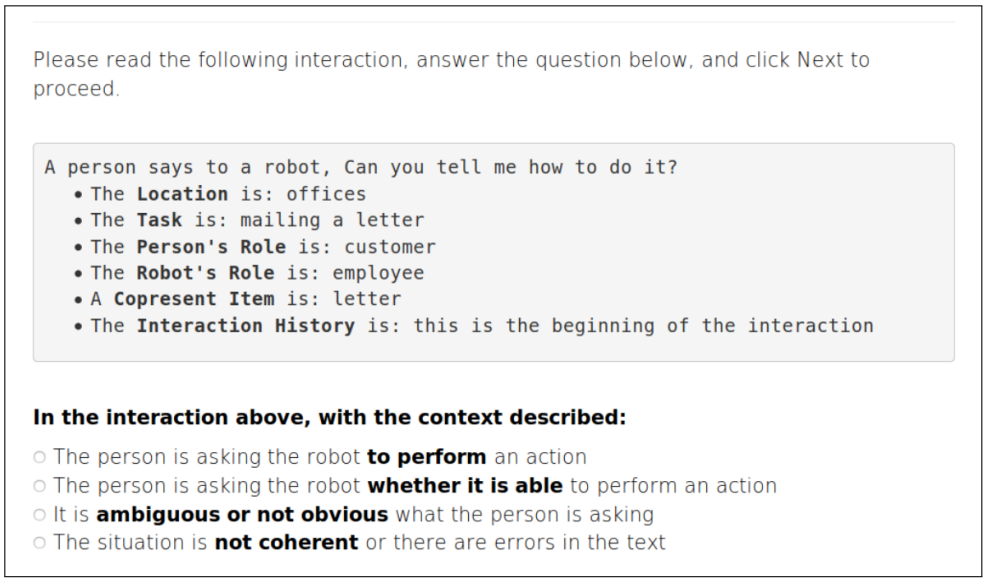

Figure 3.5: Example of the web-based GUI used for non-expert validation of ISA Schemas.

GUI to identify whether: (1) the intended meaning is to perform an action, (2) the intended meaning is a question about whether the hearer is capable of performing the action, (3) whether the half-schema is ambiguous or not obvious, or (4) whether the text is incoherent (which may be due to the nature of the automated extraction of "Can you..." questions, or due to schema creators who completed the task poorly.)

171 judgments were completed by 12 annotators perfoming 13 judgments each, plus 15 judgments by annotators who did not complete the process of authoring a full set of 13. (15 of these judgments were therefore "duplicate" judgments of schemas that had already been judged.)

The results are shown in Table 3.2. We note that non-expert validators had

achieved a score of 92% on WS, as described in Section 3.7.3 That was on expert-authored WS, and authoring WS is arguably easier than authoring $S_{ISA}s$, so we expected a fairly low accuracy rate on this, and indeed the result of 37% (63 out of 171) bears this out. We believe this low accuracy was due to both the difficulty of ISA authoring and the use of non-experts, though further work will need to be done to determine this.

Validator Response	*Total*	*Req. Action*	*Ask Ability*
Agreement	**63**	37	26
Disagreement	**76**	45	31
Ambiguous	**26**	13	13
Incoherent	**6**	1	5

Table 3.2: Results of Non-Expert Validation Study 1. Validator responses (agreement/disagreement with intended meaning, ambiguous, incoherent), total counts, and breakdowns by intended meaning (requested action and asking about ability.)

We performed a second validation study to confirm the results of the first one. 155 judgments were completed by 12 annotators performing 13 judgments each, where 1 judgment was left blank. The judgments for this validation study were made on the same schema as the first validation study. The results are shown in Table 3.3; the human accuracy rating of 44% (68 out of 155) is comparable to the first study.

Another reason to perform a second validation study is to enable inter-rater reliability: in other words, to determine the extent to which crowdsourcing annotators agree on the intention of the authored utterance.

There were 107 comparisons in which an annotator a_1 from the first validation study and an annotator a_2 from the second validation study looked at the same half-schema and determined that it was neither ambiguous nor incoherent. Of those 107 comparisons, in 30 cases a_1 and a_2 agreed with the schema author. In 12 of those cases they agreed that it was an ask-ability, and in 18 cases they agreed that it was a request-action. However, recall that these annotations are actually on half-schemas. In fact, of those 30 cases only 1 schema is formed from an appropriate set of half-

Validator Response	*Total*	*Req. Action*	*Ask Ability*
Agreements	**68**	38	30
Disagreements	**68**	36	32
Ambiguous	**13**	6	7
Incoherent	**6**	3	3

Table 3.3: Results of Non-Expert Validation Study 2. Validator responses (agreement/disagreement with intended meaning, ambiguous, incoherent), total counts, and breakdowns by intended meaning (requested action and asking about ability.)

schemas. This indicates the limits of attempting to rely completely on non-expert authoring of $S_{ISA}s$: it appears to be possible, but the yield is extremely low.

That is why for **Step 5** the expert author serves as tiebreaker for the 107 half-schemas which the validators agreed were coherent and unambiguous (but for which the validators disagreed about the intended meaning). This resulted in a total of 36 full $S_{ISA}s$.

One of the goals of this particular corpus development was to determine the strengths and shortcomings of our approach, so at this point we closely examined and manually edited the entire corpus (whose size had been limited to enable this.) In general, this step would be avoided to minimize expert authoring for scalability purposes. Alternately, a randomly-selected subset of the data could be examined to get a sense of the quality of data being produced.

Generally, the schemas produced after tiebreaking only required a few edits (2-3 of the context features per schema) to reach the level of quality that satisfied an expert author. However, some utterances were more ambiguous and therefore especially difficult for the participants to develop a schema from. For example "can you tell me about it" was difficult as most of the time the implied meaning is for the hearer to both respond about their capability and also to perform the action. This was found in most of the "can you X me Y" utterances. This suggests that utterances of this particular sub-form are best not used for schema development by non-experts. In total, 6 of the 36 utterances produced by tie-breaking were identified as ambiguous in this way, resulting in a final total of 30 $S_{ISA}s$, or 60 data points for

testing an ISA interpretation system.

[*l*]utterance	*Can you tell me something about the patent law in our country from the book you are reading?*	
literal-meaning	*ask-ability*	
context-1	[*l*]task	*asking about a book*
	speaker-role	*student*
	hearer-role	*teacher*
	location	*school*
	copresent-item	*notepad*
	interaction-history	*this is the beginning of the interaction*
intended-meaning-1	*request-action*	
context-2	[*l*]task	*getting contact information*
	speaker-role	*administrator*
	hearer-role	*student*
	location	*phone call*
	copresent-item	*phone*
	interaction-history	*they have talked about events related to the task*
intended-meaning-2	*ask-ability*	

Figure 3.6: ISA Schema authored from non-expert crowdsourcing.

3.10 Discussion and Future Work

To ensure the quality of the corpora produced, crowdsourcing with non-experts and the extraction of schema elements from corpora should not be seen as techniques that replace expert authors altogether, but instead as techniques that greatly reduce the authoring burden of experts, and also provide a basis for schemas that are grounded in reality. Consider the schema shown in Figure 3.6, which was produced by non-expert agreement alone (i.e. the two non-expert validators agreed with the non-expert author.) Even this schema would benefit from expert authoring modifications: in context 2, for example, the task could be re-stated as "trying to find the right person to talk to."

Future work includes further exploring authoring techniques to improve the effectiveness of the non-experts. There are at least two types of improvements we could consider. First, we could explore incremental changes such as refining the GUIs that the non-experts use. For example, we could revise the suggested context feature options in the drop-down menus. In Step 3 of our example corpus development, the

non-expert authors often identified one or more contexts which changed the utterance's meaning, but the non-expert authors often failed to identify those contexts which were irrelevant, and often left several "confound" context features such as co-present objects. Also, the non-expert authors often seemed to be confused about the difference between tasks and roles. Although they were generally able to identify a relevant location, the authors often did not take advantage of writing in their own feature values. Finally, the feature values could be modified. For example, we could describe roles in terms of the dynamic between people, such as "transactional" for customer-employee and "leader-follower" for manager-employee.

The second type of change we could make involves breaking up the authoring task performed by non-experts. In Step 4 of our example corpus development, we noted that we achieved a low accuracy rate on non-expert validation, and we suggested that the difficulty of ISA authoring might be a cause of this. Specifically, the non-expert authoring of $S_{ISA}s$ in Step 3 is a challenging mental task involving linguistic phenomena that most people have not extensively reflected upon. It might therefore be better to break up that authoring task into several sub-tasks that people find more intuitive. For example, Step 3a might involve presenting an utterance and asking the non-expert author to imagine a context in which that utterance had the intended meaning "request-action." If so, they would be prompted to write a brief paragraph in plain language explaining *why* the intended meaning followed from the context, specifically listing the relevant context details. Step 3b of the process would involve presenting these brief paragraphs to a second set of non-expert authors, and asking them to author the context feature/value pairs (such as "task: mailing a letter,") i.e. the $c_{1...n}$ of the $S_{ISA}s$. The advantages of this approach are: (1) the cognitive tasks of authoring the scenario and formalizing the context are separated, allowing non-expert authors to focus on one at a time, (2) Step 3b serves as an initial validation check, (3) the $S_{ISA}s$ could then also include a *meta-why-n* field explaining in plain language the scenario author's reasoning for the given interpretation in context n. The potential disadvantage of this approach is that it may involve more non-expert authors, although the total amount of time those authors take may be

lower; this can be investigated and quantified.

To summarize, we have defined an approach that balances multiple techniques to develop a corpus of $S_{ISA}s$. The utterances in Step 1 are extracted from corpora; the context features in Step 2 are authored by experts from corpora; the initial schemas in Step 3 are authored by non-experts using crowdsourcing; the validations in Step 4 are similarly performed by non-experts using crowdsourcing; in Step 5 an expert author acts as tie-breaker to produce validated schemas. The example corpus development shows that non-experts are capable of developing schema halves which can then be assembled by expert authors.

There is much possible follow-up work involving improving authoring techniques by incremental GUI performance, as well as by splitting up the authoring tasks. Finally, additional issues for further study will doubtlessly be identified when the corpus of $S_{ISA}s$ is actually used to investigate a system's Indirect Speech Act processing.

3.11 Conclusion

In this chapter, we present a novel cognitive scientific model of the reasoning involved in automatically interpreting Indirect Speech Acts (ISAs), empirically derived from a corpus analysis. The approach is grounded in existing cognitive science and psycholinguistic theories and provides a domain-general formal axiomatization of the reasoning needed to uncover the intended meanings of ISAs.

Per the approach, the listener reasons about the speaker's intentions and beliefs, using contextual evidence that both the speaker and listener have access to. Unlike previous approaches, we incorporate reasoning about a speaker's beliefs about the listeners capabilities, socio-normative standards, and speaker preferences. Moreover, extending past approaches, we provide a way to incorporate various aspects of extra-linguistic context into this reasoning process. Following traditions of building formal representations from corpora, we ensure that our representational choices completely cover utterances across several task-oriented dialogue corpora that fea-

ture ISAs. We focus on those utterances that have similar surface form, but different meanings due to contextual factors. In future work, we intend exploring how linguistic norms of directness, politeness and brevity, associated with ISAs [LW20, WSW20] can be incorporated into the reasoning process either within or as an extension to Brown's model.

We encode the formalism in a logic-based language for knowledge representation and reasoning. We use a logical representation for several reasons. First, such a representation allows for explainable processing; specifically, a system would be able to justify the "thought-process" behind a given interpretation. Such explainability allows the human user to detect mistakes in and debug the system, and helps build trust in the system [Rei19]. Second, such a representation could be used to support online learning: if a system produced a given interpretation, and it later discovered that the interpretation was incorrect, it would have an explicit representation that it could use to reason about whether the error was in its contextual evidence, its domain-general rule application, or its preconditions. Finally, we present this model as a standalone scientific contribution in its own right, building on the work of Brown, Clark, and the others cited in Section 3.3; a contribution in the fields of cognitive science and psycholinguistics.

Chapter 4

Perceiving Object Affordances

In this chapter, I turn to the role of logical sense-making in recognizing how objects can be used and manipulated, namely their affordances. Although a different representational language from the one used in language understanding (Chapters 2 and 3), the basic thesis that logical sense-making is needed for visual perception and action selection still remains. In addition to knowledge representation, here the focus is less about reasoning and about the role of uncertainty and how such sense-making works at an architectural level. I discuss the sorts of cognitive architectural components (implemented in a robotic system) might be needed in conjunction with a representational format.

4.1 Introduction

Natural human activities involve using and manipulating objects around us and continuously reasoning about our environment. Consider the example of cooking activities in a restaurant kitchen: these activities require cutting vegetables, monitoring the stove and keeping tools and utensils clean, all while ensuring orders are prepared and served in a coordinated and timely manner. Not only are team members able to recognize various objects around the kitchen, but they know what to do with these objects, how to use them appropriately, how to help others use them (i.e., they can infer and act on complex object affordances). That is, the kitchen

team is using these affordances to reason about the task at hand. Sometimes these types of activities involve standard reasoning tasks, like choosing a clean knife for cutting a tomato. Other times, these activities involve more creative reasoning tasks like solving puzzles and finding novel uses for objects, like using a dishcloth as an oven mitt.

Reasoning and using objects in this manner is a highly desirable skill for robotic agents as well. Helper robots will be critical in many application domains: helping our elderly and disabled in assisted living facilities, conducting search-and-rescue missions in unforgiving terrain to save human lives, assisting our astronauts on the space station, or even monitoring our surroundings to keep us safe from national security threats. In these critical sectors it is highly beneficial to endow robots with the ability to find creative ways to use and manipulate objects, especially when there is minimal and uncertain information. Unfortunately, although today's robots are proficient at recognizing object features, they are less skilled at recognizing what can be done with these objects.

In this work, we present a novel computational framework based on Dempster-Shafer (DS) theory [Sha76] and "uncertain logic" for inferring object affordances. Our framework comprises a logic-based representational format and inference mechanism coupled with a nascent computational architecture, CALyX (Cognitive Affordances Logically eXpressed), to reason about not only functional and physical features of objects, but also social, historical, aesthetic and ethical aspects that we naturally consider when perceiving objects – generally, "cognitive affordances". For example, we know that dirty knives are typically not used for cutting vegetables, even though they can functionally accomplish the task. As such we will demonstrate, with examples, that with our proposed approach a robot will be able to reason about these kinds of complicated affordances in a unified, systematic, and effective manner.

4.2 Cognitive Science Background

4.2.1 The Concept of Affordances

Gibson introduced the notion of "affordance" to account for human visual perception [Gib79]. He considered affordances as latent properties of the environment that exists in the presence of animals and humans. His general description captured the deeply interconnected relationship between an animal and its environment in ecological terms. Since then, the idea of affordances has been adopted across a variety of disciplines including psychology, computer science, artificial intelligence, and human-computer design. However, despite this extensive adoption, the ontological and representational aspects of affordances have been the subject of vigorous debate.

In general, while researchers agree that affordances are possibilities for action, they disagree with respect to what they are, where they are and how they work. There are numerous different ways that researchers have conceptualized affordances [Ruc20, AS19, BR83, BKR18, Che11, Den17, Gal18, RJ19, Gła15, GW12, HEHEG19, Kri18, Leu12, MCR+18, NN+04, RK14, Ruc17, Var88, VTR16, WDPAP12].

Two contested questions focused on were what affordances are and where they are supposed to live? Do they belong to the environment, to the agent, or to the agent-environment system? Turvey proposed that affordances are dispositional properties of the environment and actualized by the actions of the agent [Tur92]. Reed proposed a more radical theory stating that affordances, although disposed in the environment, are a scarce resource and actually play a role in regulating human adaptive behavior and natural selection [Ree96]. Norman proposed two different kinds of affordances: a real affordance that is in the environment and a perceived affordance that is in the agent's mind [Nor88].

Stoffregen argued that affordances do not belong to either the agent or the environment, but are instead emergent properties of an agent-environment system [Sto03]. Several theories also explored explicit representational formats including describing affordances as relations connecting the abilities of the agent with environmental features [Che03], and further connecting the effects of agent behaviors on

the course of events [SCD+07], or as relations connecting environmental attributes in overlapping conceptual spaces or regions [MSSZ11, MS12]. Scarantino claimed that affordances are not only relational, but also conditional [Sca03]. Specifically, she argued that affordances are conditional upon various triggering conditions and related to the agent's set of potential abilities.

A number of these and other theories focused primarily on functional aspects of affordances [BSC02, BA03]. There has also been some limited work in introducing social considerations into an affordance framework. Schmidt argued to extend Scarantino's theory of conditional and relative nature of affordance to include the idea of social affordances [Sch07]. Work by Kim in the particular space of cognitive robotics and object handover considered social etiquette and norms as well [KP04, SLD+13].

Using and manipulating objects involves not only functional and physical aspects of objects, but other features including social conventions that govern the object's use, aesthetic considerations that limit what can and cannot be done with an object, ethical factors that guide moral action, and historical precedence that influences the designed purpose and intent for the objects. An affordance inference framework must allow for a broad definition of affordance, which we refer to as "cognitive affordance", one that accounts for functional as well as non-functional aspects and must be adaptable to allow for continuous changes to and evolution of these aspects over time. Some of the above-mentioned theories and representations are limited in their ability to reason about affordances more holistically and contextually.

In the next section, we will describe past approaches to reasoning with affordances as used in cognitive robotics. These approaches adopt some of the more conceptual theories mentioned above, e.g., those by [SCD+07, MS12, KP04, SLD+13] and implement them in computational and robotic systems. Discussing these approaches allows us to more specifically place our own contribution in the context of past work.

4.2.2 Affordances in Cognitive Robotics

In cognitive robotics, there have primarily been two types of approaches to representing, inferring, and reasoning with affordances: (1) approaches based on statistical and machine learning formalisms, and (2) approaches based on ontological formalisms. These are very powerful approaches and have shown substantial benefits to robotic cognition. However, as we will discuss, these approaches are limited both representationally and architecturally. Specifically, they do not demonstrate flexible representational formats to account for social and other non-functional aspects of affordances, they do not allow for contextual reasoning, and they do not address uncertainty in perception and beliefs. These approaches are also limited architecturally because they mostly only involve bottom-up processing of sensory (mostly visual) information and thus do not allow for much top-down processing of sensory information, which is necessary for a more complete account of affordance perception.

Steedman used Linear Dynamic Event Calculus to formalize the relationship between objects and their affordances [Ste02]. More recently, work by Abel and Tellex focused on using Markov Decision Processes to directly model affordances as mappings between a set of preconditions and goal states, to action possibilities [ABMMT15]. Mastrogiovanni et al. have developed a framework, using Self-Organizing Neural Maps, for action selection and functional representation of everyday objects, places and actions in terms of affordances and capabilities, as regions in a proper metric space [MSSZ11, MS12].

The strength of these works lies in their joint modeling of affordances with the problem of planning and action sequencing. It allows for not just reasoning about actions, but also implementing action sequences that then allow for new affordances to emerge. However, while affordance perception involves action selection from a choice of action capabilities, its inference has broader applicability than just for planning. Affordance inference is important to other cognitive processes involved in commonsense reasoning, natural language explanations, and general environmental

sense-making. We believe there is a benefit to representing and reasoning with affordances in a manner that disentangles it from planning, but still allowing for leveraging the extensive advances in the planning literature.

Montesano et al. have developed statistically-inspired causal models of affordance using Bayesian networks to formalize the relationship between object features, actions, and effects [MLBSV07, ML09]. Several others have modeled affordances as a relationship between action, object, and effect [UCS+11, USO12, UNO13, MMV+12]. A number of computational and robotic systems have also emerged to tackle various sub-problems relating to robotic affordances such as object grasping and handover [AMC14, CPCI15, UNO13].

The strength of these works lies in their underlying model of affordances per Sahin's approach of relating objects, actions, and the effects [SCD+07], allowing for a close relationship with planning. But here too, inference of affordances is not separate from specific planning tasks and, therefore, is not applied more generally.

A few researchers have explored ontology-based approaches to represent functional affordances. For example, Varadarajan et al. have developed a detailed knowledge-ontology based on conceptual, functional and part properties of objects, and then used a combination of detection and query matching algorithms to pinpoint the affordances for objects [VV11, Var15]. While being able to query an affordance knowledge-base is helpful from a deductive standpoint, this approach is limited in its flexibility for accounting for contextual shifts, and changing social norms.

Moreover, the focus on much of the affordance work in cognitive robotics is on functional affordances, and so there is often no distinction provided between a hammer in a person's toolbox and a decorative hammer on display at the museum, both of which are functionally equivalent, but engender entirely different non-functional affordances. The social affordances associated with interacting with a museum object are vastly different from the social affordances of interacting with a hammer in a personal toolbox.

Shu et al. have recently presented a framework for reasoning about social affordances and provide a system that can act in social scenarios like handshaking,

helping a person stand up, high-fiving, and handing over objects [SRZ16]. While Shu is reasoning about affordances in social interactions, the underlying affordance model is still largely devoid of contextual reasoning, and focused more on physical geometries of objects in these scenarios (in this case skeletal geometries). However, such physical aspects do not account for the contextual information that is not perceptual (e.g., high-fiving a friend versus a refraining from high-fiving an enemy) and is also subject to change.

Thus more generally, despite these past efforts, affordance representation faces many challenges that have not been overcome in the previous work. Specifically, past approaches fail to provide flexibility with which to reason about affordances in the open world, where they are influenced by changing context, social norms, historical precedence, and uncertainty. For example, none of the current approaches can systematically infer that coffee mugs afford grasping and drinking, while also simultaneously affording serving as a paperweight or cupholder, or depending on the context, as family heirloom not meant to be used at all. We argue that inferences of this sort are different from sole high-level reasoning or planning processes, for they require a continuous interplay between low-level sensory systems and high-level cognitive systems and between bottom-up (sensory mechanisms to higher-level cognition) and top-down processing (higher-level cognition to sensory mechanisms) of information in these systems. Critically, cognitive affordance representation and reasoning is a separate cognitive process in its own right and deserves its own architectural framework and inference machinery (separate from high-level reasoning and planning or low-level feature detection) that can then later be tied together with suitable perceptual and planning and reasoning frameworks. This is not to say that affordances are not influenced by perception, planning, and reasoning – they are – but affordance-based reasoning is not fully explained by and thus not subsumed within these processes.

Next, we present an architecture for reasoning about affordances that has components distinct from perceptual processes (e.g., vision, haptics) and from action processes (e.g., planning and natural language interaction). Our framework

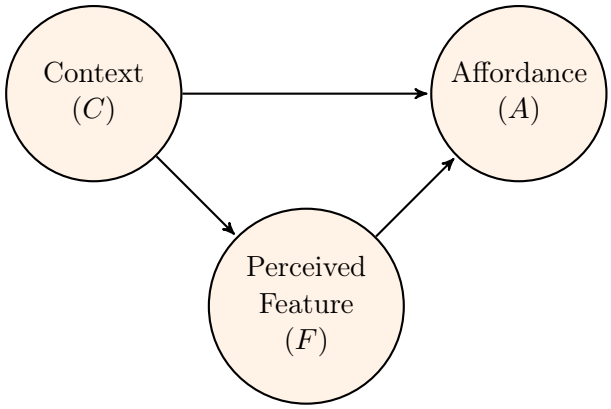

Figure 4.1: Context-Sensitive Cognitive Affordance Model

enables reasoning about higher-level affordances that rely on cognition and contextual reasoning separately from perception and action.

4.3 The Computational Cognitive Affordance Framework

The proposed computational cognitive affordance framework consists of (1) a logic-based affordance representation and (2) a computational architecture (CALyX) that is context-sensitive and furthermore allows for top-down constraints on visual perception of the environment. Note that the proposed CALyX architecture is distinct from a (low-level) vision systems even though affordance reasoning can interface with it. Rather, the affordance representations used in CALyX are agnostic to the originating modality of the percepts (e.g., vision, haptics, natural language, etc.), allowing for reasoning at a higher than sensory-level (the sensory-level is sometimes referred to as detection in ecological psychology). Different from mere sensory processing of affordances, the higher-level representations and reasoning processes take into account perceptual, task-based, and other context as well as relevant mental states of the agent such as beliefs, intentions, goals, and desires.

4.3.1 Logic-Based Representation

We propose a novel representational format for cognitive affordances, illustrated in Fig. 4.1, in which an object's affordance (A) and its perceived features (F) depend

on the context (C) [SS15a]. We use Dempster-Shafer (DS) theory [Sha76] – an uncertainty processing framework often interpreted as a generalization of the Bayesian framework – for inferring affordance (A) from object features (F) in contexts (C). More specifically, the proposed cognitive affordance model consists of four parts: (1) a set of perceivable object features (F), (2) a set of context states (C), (3) a set of object affordances (A), and (4) a set of "affordance rules" (R) connecting object features and context states to applicable affordances which take the overall form:

$$r :\equiv f \wedge c \implies_{[\alpha,\beta]} a$$

with $f \in F$, $c \in C$, $a \in A$, $r \in R$, $[\alpha, \beta] \subseteq [0, 1]$. Here, the confidence interval $[\alpha, \beta]$ is intended to capture the uncertainty associated with the affordance rule r such that if $\alpha = \beta = 1$ the rule is logically true, while $\alpha = 0$ and $\beta = 1$ assign maximum uncertainty to the rule. Rules can then be applied for a given feature percept f in given context c to obtain the implied affordance a under uncertainty about f, c, and the extent to which they imply the presence of a.

We have previously shown that these types of rules are very versatile and that we can employ "DS-theoretic modus ponens" to make uncertain deductive and abductive inferences [WBOS15a]. Most critically, these rules allow us to address representational challenges with Bayesian models where $P(A|F, C)$ needs to be inferred by way of $P(F|A, C)$, $P(A|C)$, and $P(C)$ when we often have no practical way of obtaining the necessary probability distributions for all the affordances for an object. We will next provide an overview of our proposed computational architecture, which we will then use in combination with the above mathematical model to reason through two situations, each involving a tight interplay between social and functional affordances.

4.3.2 Computational Architecture (CALyX) - Overview

4.3.2.1 Introduction

We now present the computational *Cognitive Affordances Logically eXpressed* (CALyX) architecture (Fig. 4.2) for perceiving and reasoning about cognitive affor-

dances in a unified manner. CALyX has two main components: (1) an *Affordance Reasoning Component* (ARC) for performing logic-based inferences of cognitive affordances, and (2) a *Perceptual Semantics and Attention Control Component* (PAC) for directing perception in a top-down manner and semantically analyzing perceptual information in a bottom-up manner. In addition, CALyX has two supporting memories: *Long-term Memory* (LTM) and *Working Memory* (WM), for storing and updating logical affordance rules and related uncertainties. These components work closely with sensory and perceptual systems (e.g., vision) and other components in a cognitive architecture to coordinate perceptual and action processing.

We will focus on the main components noted above and briefly touch upon other cognitive components as and when needed. It is important to note here that CALyX is only a part of a larger cognitive architecture and as such we do not expect it to cover other cognitive subsystems (e.g., those for planning or natural language processing) or provide an account for all manner of cognitive function. Instead, we focus on affordance perception and inference and note that CALyX serves as a intermediary subsystem linking lower level perceptual subsystems (e.g., vision, motor control, haptics) with higher-level belief, planning and goal management systems to facilitate top-down and bottom-up processing in the larger cognitive architecture (e.g., the DIARC architecture within which CALyX was developed [SSKA07]).

4.3.2.2 Cognitive Cycle

In each cognitive cycle, ARC selects applicable rules from LTM and populates WM. Once the rules are in Working Memory, both PAC as well as ARC use these rules as the basis for perception and inference. More specifically, PAC directs low-level perceptual systems like vision to perform visual searches in a focused manner only looking to determine beliefs for the specific perceptual features, F, relevant to the applicable rules in Working Memory. This is a top-down attentional strategy that helps the robot focus its senses on relevant parts of the environment given the rules in the WM while ignoring others. ARC performs DS-theoretic affordance inference on the rules in WM using beliefs about the relevant perceptual features from PAC

and beliefs about contexts provided by other parts of the cognitive architecture. The outcome of the inference process is the generation of truth values of various affordances specified in WM and their associated uncertainty intervals, which are then used by the rest of the cognitive architecture for planning, reasoning and sense-making tasks.

4.3.2.3 Memory Management and Context-Sensitivity

In any given situation the robot might be subject to a set of overlapping contexts. For example, in a situation in which a robot is a kitchen helper, it might be subject to a context that refers to its role as a helper-robot. Simultaneously, the robot may also be in a more specific context that refers to a particular task that it must perform, for example, the task of handing over a knife to the human chef. This set of contextual aspects, C, collectively constitutes the agent's situation. As noted earlier, contexts may often not be perceivable, containing non-perceivable aspects of the task context and environment as well as the agent's own belief system. The fact that the agent is a kitchen helper, for example, is not necessarily perceivable by simply visually scanning the environment. Information about the robot's role, beliefs, desires and intentions may be provided by other high-level processing components in the robot's cognitive architecture. Thus, contextual aspects represent those descriptors of the situation which can be non-physical and even abstract.

This contextual information is passed into CALyX from other parts of the cognitive architecture and received by a Memory Management subcomponent of ARC. The Memory Management subcomponent searches through all available affordance rules of the form specified above in the agent's LTM and identifies rules that contain contexts that match those in the current situation. We use a matching threshold ζ to determine whether the current context "matches" the context presented in the rule. We can set the mass threshold to a value $0 \leq \zeta \leq 1$ and check if the mass of the current context exceeds this ζ threshold. Contextual aspects with masses satisfying the threshold condition will be considered by the Memory Management subcomponent. The Memory Management subcomponent aggregates the applicable rules (i.e., rules

applicable to contexts in the current situation) and populates WM. The WM stores rules of the form described earlier along with corresponding uncertainty intervals.

The Memory Management subcomponent passes the contextual aspects along with their mass assignments to the Affordance Inference subcomponent. The Affordance Inference subcomponent also has access to the WM of rules and accompanying uncertainties. In order to perform inference, the Affordance Inference subcomponent also needs uncertainty information about the set of perceptual aspects, F, identified in the applicable rules stored in WM. For this, it will turn to PAC.

4.3.2.4 Attention Control and Perceptual Semantics

PAC accesses the rules in WM and determines what perceptual aspects need to be evaluated. For example, if PAC needs to compute if there are grasp locations near the handle of a knife, it can resolve this perceptual relation query:

$$near(knife, G, holdPart(knife) = handle)?$$

PAC includes vision algorithms to resolve various sorts of relations including the spatial relation of $near()$. PAC directs low-level perception subcomponents (e.g., vision) to look for and identify uncertainties associated with relevant perceptual aspects. PAC returns to ARC the masses associated with perceptual aspects of the applicable rules.

4.3.2.5 Affordance Inference

The Affordance Inference subcomponent of ARC then performs DS-theoretic inference on the rules in WM using masses for the contextual aspects obtained from the Memory Management subcomponent and masses for the perceptual aspects obtained from PAC. ARC computes the uncertainties associated with affordances prescribed by the rules. In certain cases, the Memory Management subcomponent will selectively populate WM with rules that not only satisfy context, but also specify relevant affordance relations. This set of rules would be a subset of the applicable rules for

the selected context.

Generally, affordance aspects and their associated uncertainty intervals and confidence measures are passed from ARC to other parts of the cognitive architecture including those subsystems responsible for planning, reasoning and sense-making.

4.4 Robot Kitchen Helper Experiment: Using and Handing Over Objects

For the experimental evaluation of the proposed computational cognitive affordance framework we will consider using and handing over objects in a kitchen as a running example to discuss the representation format, the uncertainty processing framework, and the inference algorithm in the implemented CALyX architecture. We will show how our framework assists the agent in reasoning about and deciding what action possibilities are available during each phase of the handover process, from grasping the object, to handing it over.

Note that handing over objects "properly" is an important skill for helper robotic agents. When performing a handover, the robot will need to reason about potential actions it can perform on objects (affordances), for example, selecting grasps or manipulating the object in certain ways. Existing approaches have focused on selecting one or two handover norms (e.g., orienting a handle towards the receiver), *a priori*, and then building object recognition and motion planning systems that are dependent on the preselected norms [AMC14, CPCI15]. These approaches fail to provide flexibility with which to reason about action choices in an open world, where norms and rules may change, norms may be added and removed, normative conflicts may arise, and other contextual factors may influence the propriety of a handover. In contrast, we intend to infer affordances based on (1) the semantic representation of certain visual percepts, (2) the agent's current context, and (3) the general domain and commonsense knowledge of the agent.

We will first provide a brief review of Dempster-Shafer theory and then use our framework to model the domain of cooking and assisting humans in the kitchen.

Then we will walk through how an agent, staffed as a kitchen helper, reasons through the process of handing over knives.

4.4.1 Mathematical Preliminaries

4.4.1.1 Dempster-Shafer Theory

A set of elementary events of interest is called *Frame of Discernment* (FoD). The FoD is a finite set of mutually exclusive events $\Theta = \theta_1, ..., \theta_N$. The power set of Θ is denoted by $2^\Theta = \{A : A \subseteq \Theta\}$ [Sha76].

Each set $A \subseteq \Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping $m_\Theta(\cdot) : 2^\Theta \to [0, 1]$ such that $\sum_{A\subseteq\Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set $\mathcal{F}_\Theta$. The triple $\mathcal{E} = \{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$ is called the *Body of Evidence* (BoE). For ease of reading, we sometimes omit $\mathcal{F}_\Theta$ when referencing the BoE.

Given a BoE $\{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B\subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) = 1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict A. The *uncertainty* interval of A is $[Bel(A), Pl(A)]$, which contains the true probability $P(A)$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$.

Dempster-Shafer theory can be considered a generalization of Bayesian theory. For example, a Bayesian would model Schrödinger's cat as a probability distribution over $\{dead, alive\}$, assigning a probability to each hypothesis. Dempster-Shafer would assign masses to each of $\{dead, alive, \{dead\ or\ alive\}\}$, without beliefs having to sum up, for example $Bel(dead) + Bel(alive) \neq Bel(dead \vee alive)$. One notable advantage of this uncertainty processing framework is that it allows for the allocation of probability masses to sets of hypotheses, and does not require an assumption about the probability distribution among members of that set.

4.4.1.2 Uncertain Logic

Logical inference with uncertainty can be performed using DS-theoretic Logical inference with uncertainty can be performed using DS-theoretic "Modus Ponens" (denoted $\odot$) as discussed by Tang et al. [THPS12]. We will use Tang's DS-theoretic AND (denoted $\otimes$) to combine BoEs on different FoDs [THPS12]. We choose to use Tang's models of Modus Ponens and AND over other proposed models because those models do not allow uncertainty to be multiplicatively combined.

We will use Yager's rule of combination (denoted $\bigcap$) to combine BoEs on the same FoD [Yag87]. Yager's rule of combination aggregates evidences or confidence values from different sources, but within the same frame of discernment. Formally, when combining evidence from n different sources within the same frame, Θ, the combined multi-evidence BBA, according to Yager's rule is defined as follows:

$$m_\Theta(\emptyset) = 0$$

$$m_\Theta(A) = \sum_{\cap B_i = A} \prod_{i=1}^{n} m_{\Theta i}(B_i), \forall A \subseteq \Theta, A \neq \Theta, A \neq \emptyset$$

$$m_\Theta(\Theta) = \prod_{i=1}^{n} m_{\Theta i}(B_i) + \sum_{\cap B_i = \emptyset} \prod_{i=1}^{n} m_{\Theta i}(B_i)$$

Yager's rule of combination is chosen because it allows uncertainty to be pooled in the universal set, and due to the counter-intuitive results produced by Dempster's rule of combination, as discussed in [Zad79].

For two logical formulae ϕ_1 (with $Bel(\phi_1) = \alpha_1$ and $Pl(\phi_1) = \beta_1$) and ϕ_2 (with $Bel(\phi_2) = \alpha_2$ and $Pl(\phi_2) = \beta_2$, applying logical AND yields $\phi_1 \otimes \phi_2 = \phi_3$ with $Bel(\phi_3) = \alpha_1 \cdot \alpha_2$ and $Pl(\phi_3) = \beta_1 \cdot \beta_2$.

For logical formulae ϕ_1 (with Bel$(\phi_1) = \alpha_1$ and $Pl(\phi_1) = \beta_1$) and $\phi_{\phi_1 \to \phi_2}$ (with $Bel(\phi_{\phi_1 \to \phi_2}) = \alpha_R$ and $Pl(\phi_{\phi_1 \to \phi_2}) = \beta_R$, the corresponding model of Modus Ponens is $\phi_1 \odot \phi_{\phi_1 \to \phi_2} = \phi_2$ with $Bel(\phi_2) = \alpha_1 \cdot \alpha_R$ and $Pl(\phi_2) = 1 - ((1 - Pl(\beta_1)) \cdot (1 - Pl(\beta_R)))$.

Moreover, we will use the "confidence measure" λ (defined in [NDS+13]) to

be able to compare uncertainties associated with formulas ϕ and their respective uncertainty intervals $[\alpha, \beta]$:

$$\lambda(\alpha, \beta) = 1 + \frac{\beta}{\gamma} log_2 \frac{\beta}{\gamma} + \frac{1-\alpha}{\gamma} log_2 \frac{1-\alpha}{\gamma}$$

$$\text{where } \gamma = 1 + \beta - \alpha.$$

Here, ϕ is deemed more ambiguous as $\lambda(\alpha, \beta) \to 0$.

4.4.2 Semantic Representation of Visual Perception, F

The vision pipeline for an artificial agent involves various low-level components that are coupled together to process color and depth information and generate point clouds and 3D meshes. As noted earlier, PAC is configured to perform scene representation and semantic analysis to generate predicates that capture, qualitatively, certain aspects of the visual scene.

Let $F = \{\Theta_{F_1}, \Theta_{F_2}, \ldots, \Theta_{F_N}\}$ be the set of N different perceptual aspects such as color, shape, texture, relational information, and generally information obtained from the vision pipeline that an agent may interpret. Each aspect $\Theta_{F_i} = \{f_{i,1}, f_{i,2}, \ldots, f_{i,M_i}\}$ has a set of M_i mutually-exclusive candidate perceptual values (percepts), which come from the vision system as a BoE, $\mathcal{E}_{F_i} = \{\Theta_{F_i}, m_{\Theta_{F_i}}(\cdot)\}$. We will use $m_{f_{i,j}}$ to denote the candidate mass values of the percepts, where $i \in \{1 \ldots N\}$ and $j \in \{1 \ldots M_i\}$.

For the purposes of our example, we will represent the agent's visual perception of kitchen objects with nine binary ($|M_i| = 2$) visual aspects, each aspect with a percept and its negation. Thus, $\Theta_{F_i} = \{f_{i,j}, \neg f_{i,j}\}$, where $i \in \{1 \ldots 9\}$ and $j \in \{1\}$. The percepts and masses for each of the nine aspects, which can be obtained from the low-level vision system, are shown below:

holdPart(O) and *funcPart*(O) are functions that return the name of the holding and functional parts of an object O. Thus, *funcPart*(*knife*) = *blade* represents the knowledge that the blade is the functional part of the knife. Similarly,

$holdPart(knife) = handle$ represents the knowledge that the handle is the holding part of the knife.

Aspect (Θ_{F_i})	Percept ($f_{i,j}$)	Mass ($m_{f_{i,j}}$)
Θ_{F_1}	$holdPart(O)$	$m_{f_{1,1}}$
Θ_{F_2}	$funcPart(O)$	$m_{f_{2,1}}$
Θ_{F_3}	$hasSharpEdge(O)$	$m_{f_{3,1}}$
Θ_{F_4}	$hasPointyTip(O)$	$m_{f_{4,1}}$
Θ_{F_5}	$hasOpening(O)$	$m_{f_{5,1}}$
Θ_{F_6}	$near(O, G,$part$)$	$m_{f_{6,1}}$
Θ_{F_7}	$grasped(O,$part$)$	$m_{f_{7,1}}$
Θ_{F_8}	$dirty(O)$	$m_{f_{8,1}}$
Θ_{F_9}	$inUse(O, H)$	$m_{f_{9,1}}$

$hasSharpEdge(O), hasPointyTip(O)$ and $hasOpening(O)$ represent the perception of various physical features on object O. In the case of knife we use algorithms developed by [SSCO08] to extract shape feature information from the object using object meshes. We then segment the objects (handle and blade) based on their relative sharpness.

$near(O, G,$part$)$ represents the location of a set of graspable points G on an object O in relation to a certain object part (holding or functional part). Thus, $near(knife, G, holdPart(knife) = handle)$ states that there are grasp points near the handle. The grasp points may be extracted from visual point clouds using algorithms developed by [TP14] that identify antipodal grasp information based on object geometries. We can then group these grasp points based on their location and proximity to the shape features noted above.

$dirty(O)$ represents a measure for whether a certain object is dirty or contains food particles. Thus, $dirty(knife)$ describes the knowledge that the knife is dirty. The value of this predicate is obtained from low-level vision components tasked with monitoring image characteristics of color and homogeneity.

$grasped(O,$part$)$ represents the agent's knowledge that it has grasped a cer-

tain part of the object. For example, $grasped(knife, holdPart(knife) = handle)$ represents the knowledge that the agent has grasped the handle.

$inUse(O, H)$ represents the agent's observation that an object O is currently in use by a person or agent H.

We selected these particular visual aspects because of their significance to the rules that we will discuss in more detail in the below sections. There is a potentially huge number of semantic aspects and relations in the environment and it would not be possible for the agent to keep track of them all. Our approach simplifies the task for PAC and the vision system to only look for certain relevant perceptual features based on the agent's current context. We envision that our set of perceptual aspects, F, may change dynamically to include and exclude percepts as contexts and situations change over time.

4.4.3 Relevant Contextual Items, C

Knowledge of the agent's current context is provided to CALyX by certain high-level processing components such as the agent's belief, planning, and goal management systems. The context is representative of the agent's beliefs, goals, desires, and intentions, along with certain other abstract constructs in the agent's situation. Together, these contextual items, processed as predicates, represent qualitatively the agent's abstract context, i.e., knowledge not directly perceivable.

Let $C = \{\Theta_{C_1}, \Theta_{C_2}, \ldots, \Theta_{C_N}\}$ be the set of all contextual aspects an agent may need to interpret. Each contextual aspect $\Theta_{C_i} = \{c_{i,1}, c_{i,2}, \ldots, c_{i,M}\}$ has M mutually-exclusive candidate contextual states, which come from the high-level components as a BoE, $\mathcal{E}_{C_i} = \{\Theta_{C_i}, m_{\Theta_{C_i}}(\cdot)\}$. We will use $m_{c_{i,j}}$ to denote the candidate mass values of the contexts, where $i \in \{1 \ldots N\}$ and $j \in \{1 \ldots M\}$.

For the purposes of our example, similar to our representation of perceptual aspects, we will represent the agent's contextual knowledge with two binary contextual aspects. The first contextual aspect represents the agent's current domain or setting, L, and it includes a contextual value (context) of being a kitchen helper and its negation: $\Theta_{C_1} = \{c_{1,1}, \neg c_{1,1}\}$. The second contextual aspect represents the

agent's tasks in the kitchen while playing two different social roles: (1) as a primary actor using objects, and (2) as a supporting assistant giving objects to others. This aspect includes two contextual values: $\Theta_{C_2} = \{c_{2,1}, c_{2,2}\}$. The contexts and masses for each of the two aspects, can be obtained from the agent's belief and planning systems:

Aspect (Θ_{C_i})	Context($c_{i,j}$)	Mass ($m_{c_{i,j}}$)
Θ_{C_1}	$domain(X, L)$	$m_{c_{1,1}}$
Θ_{C_2}	$task(X, use, O)$	$m_{c_{2,1}}$
	$task(X, give, O)$	$m_{c_{2,2}}$

$domain(X, L)$ represents the agent's, X, current domain, L. For example, $domain(self, kitchen)$ represents the knowledge that the agent is currently in the domain of working in the kitchen. The reason for the domain context is to help the agent constrain the set of possible affordances available on the object to the domain it is currently in. For example, the agent might not need to consider affordances of a knife as a camping tool or as a self-defense tool, while it is functioning as a kitchen helper. Thus, by choosing a domain, we can restrict what types of affordances the agent needs to reason about in its current task. This is not to say that the agent cannot think creatively or absorb affordance rules from other domains. But, as a simplification for this example, we choose contextual aspects that can help the agent effectively manage the computational complexity of affordance inference.

$task(X, use, O)$ represents the agent's, X, understanding of its current task-related context as being that of "using" object O. For example, $task(self, use, knife)$ means that the current task-context is that of the agent using the knife for its intended purpose of cutting.

$task(X, give, O)$ represents the agent's, X, understanding of it's current task-related context as being that of "giving" or "handing over" object O. For example, $task(self, give, knife)$ means that the current context is that of the agent handing over the knife to another.

We will discuss the rules themselves in more detail in the next sections.

4.4.4 Cognitive Affordances, A

The next part of the representational framework are the cognitive affordances A computed by CALyX based on applicable rules in WM. We use affordances here to represent action possibilities available to the agent at any given moment in time. The affordances are represented semantically with predicates for action possibilities.

Let $A = \{\Theta_{A_1}, \Theta_{A_2}, \ldots, \Theta_{A_N}\}$ be the set of N different cognitive affordance aspects. Each aspect $\Theta_{A_i} = \{a_{i,1}, a_{i,2}, \ldots, a_{i,M}\}$ has a set of M mutually-exclusive candidate affordance values (affordances), which come as a BoE, $\mathcal{E}_{A_i} = \{\Theta_{A_i}, m_{\Theta_{A_i}}(\cdot)\}$. We will use $m_{a_{i,j}}$ to denote the candidate mass values of the contexts, where $i \in \{1 \ldots N\}$ and $j \in \{1 \ldots M\}$.

For the purposes of our example, we will represent the agent's affordances with eight binary affordance aspects, each aspect with an affordance and its negation. Thus, $\Theta_{A_i} = \{a_{i,j}, \neg a_{i,j}\}$, where $i \in \{1 \ldots 8\}$ and $j \in \{1\}$. The percepts and masses for each of the eight aspects, can be obtained from our rules:

Aspect (Θ_{A_i})	Affordance ($a_{i,j}$)	Mass ($m_{a_{i,j}}$)
Θ_{A_1}	$cutWith(X, O)$	$m_{a_{1,1}}$
Θ_{A_2}	$pierceWith(X, O)$	$m_{a_{2,1}}$
Θ_{A_3}	$containWith(X, O)$	$m_{a_{3,1}}$
Θ_{A_4}	$graspable(X, O,\text{part})$	$m_{a_{4,1}}$
Θ_{A_5}	$sanitizeable(X, O)$	$m_{a_{5,1}}$
Θ_{A_6}	$useable(X, O)$	$m_{a_{6,1}}$
Θ_{A_7}	$giveable(X, O, H)$	$m_{a_{7,1}}$
Θ_{A_8}	$setOnTable(X, O, U)$	$m_{a_{8,1}}$

We will discuss each of these affordance aspects below:

4.4.4.1 Commonsense Physical Affordances

Various objects in the kitchen like knives, forks, spoons pots, pans, and appliances offer the agent with various physical affordances. Here we will consider three such affordances offered by a number of different objects: (1) $cutWith(X, O)$, (2)

$pierceWith(X, O)$ and (3) $containWith(X, O)$, each representing an affordance of an object O available to an agent X in a kitchen scenario. Objects can have none of these affordances or one or more of them. For example, knife can have the affordance of cutting as well as piercing, depending on the shape of the knife.

4.4.4.2 Grasp Affordances

Many objects in the kitchen tend to have a use for which they are designed, and accordingly allow for holding and using the object in a particular way for this intended purpose. For example, knives are designed for cutting and thus can be grasped by the handle and used to cut with the blade. We account for grasp affordances with a $graspable(X, O$,part) predicate, which represents that the object O is graspable by agent X at a certain part of the object. Thus, $graspable(self, knife, holdPart(knife) = handle)$ represents that the knife's handle has a grasp affordance in the current context.

4.4.4.3 Social Affordances

In the context of a kitchen, there are a number of social norms and rules that apply to ensure safety, etiquette, cleanliness and a generally friendly atmosphere. These rules present social affordances, i.e., action possibilities related to social interaction that can be made available to the agent. Here we consider $sanitizeable(X, O)$, which represents the possibility of washing and cleaning an object. As we will see with respect to the rules in the next section, social affordances can be represented both explicitly, as in $sanitizeable(X, O)$, and implicitly via socially-derived rules for conduct, e.g., presenting the handle first when giving objects to others.

4.4.4.4 Object Manipulation Affordances

Once the agent has begun interacting with the object, certain new affordances are made available to the agent: $useable(X, O)$ represents the agent's X ability to use object O for its intended purpose; $giveable(X, O, H)$ represents the agent's X ability

to give object O to a human or another agent, H; and $setOnTable(X, O, U)$ represents the agent's X ability to place object O on surface U. These affordances allow the agent to consider its action possibilities once it is in the possession of the object.

Now, we recognize that these affordance are always available to the agent: the agent can cut, grasp, give, wash and place the knife at any time. Our affordance representation does not deny that latent affordances may exist in objects, but merely attaches uncertainties to their potential applicability. Certain dormant affordances will have low uncertainties unless certain contextual situations arise, and our rules seek to capture this type of reasoning with affordances.

It could also be argued that there are many more affordances for knives, and that we are limited in considering only a few. We agree with this argument and only present this exemplary set for demonstration and evaluation purposes. In reality, there are many more affordances, possibly unlimited, and our cognitive affordance inference framework can reason about all of them simultaneously. Although we will not address the issue of whether or not there are infinitely many affordances, we will contend that it suffices to consider only a finite number of them in any given set of contexts, applicable at a particular moment in time.

4.4.5 Cognitive Affordance Rules, R

The fourth part of our representational framework is the set of rules, R, that represent the cognitive affordance aspects, A, of the perceptual aspects, F, in contextual aspects, C. We will present an exemplary set R of rules for the handover example below.

Let $R = \{\Theta_{R_1}, \Theta_{R_2}, \ldots, \Theta_{R_N}\}$ be the set of N different cognitive affordance rule aspects. Each rule aspect $\Theta_{R_i} = \{r_{i,1}, r_{i,2}, \ldots, r_{i,M}\}$ has a set of M mutually-exclusive candidate rule values (rules), which come as a BoE, $\mathcal{E}_{R_i} = \{\Theta_{R_i}, m_{\Theta_{R_i}}(\cdot)\}$. We will use $m_{r_{i,j}}$ to denote the candidate mass values of the contexts, where $i \in \{1 \ldots N\}$ and $j \in \{1 \ldots M\}$.

For the purposes of our example, we will represent the agent's affordances with 18 rule aspects (representing 18 rules), each aspect with a rule and its negation.

Thus, $\Theta_{R_i} = \{r_{i,j}, \neg r_{i,j}\}$, where $i \in \{1 \ldots 18\}$ and $j \in \{1\}$. The percepts and masses for each of the 18 aspects, can be obtained from our rules:

Generally, the rules are of the form:

$$r^{i,j}_{m_{fc \to a}} := f \wedge c \implies a$$

The belief function, $Bel(R)$, captures the total support that can be committed to a rule, R, without also committing to the negation of the rule. The plausibility of R is, $Pl(R)$, corresponds to the total belief that does not contradict R. Together, the belief and plausibility represent the uncertainty interval, $[\alpha = Bel(R), \beta = Pl(R)]$. Thus, we write the rules in the form:

$$r^{i,j}_{[\alpha_{i,j}, \beta_{i,j}]} := f \wedge c \implies a$$

Below, we show each of the 18 rules for this example, presenting the uncertainty intervals for each of the rules. We have chosen uncertainty intervals in such a way that the more specific the rule, the more certainty and higher degree of belief the agent has about that particular rule. Thus, more specific the rule, narrower the uncertainty interval and higher the values for α and β. Also, for ease of reading, we have omitted the index $j = 1$.

Commonsense Physical Rules:

$r^1_{[0.8,1]} := hasSharpEdge(O) \wedge$
$domain(X, kitchen) \implies$
$cutWith(X, O)$

$r^2_{[0.8,1]} := hasPointyTip(O) \wedge$
$domain(X, kitchen) \implies$
$pierceWith(X, O)$

$r^3_{[0.8,1]} := hasOpening(O) \wedge$

$domain(X, kitchen) \implies$

$containWith(X, O)$

General Social Rules:

$r^{4}_{[0.95,0.95]} := dirty(O) \wedge$

$domain(X, kitchen) \implies$

$sanitizeable(X, O)$

$r^{5}_{[0.95,0.95]} := \neg inUse(O, H) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, holdPart(O))$

$r^{6}_{[0.95,0.95]} := \neg inUse(O, H) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, funcPart(O))$

General Object Grasp Rules:

$r^{7}_{[0.55,0.95]} := near(O, G, holdPart(O)) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, holdPart(O))$

$r^{8}_{[0.55,0.95]} := \neg near(O, G, holdPart(O)) \wedge$

$near(O, G, funcPart(O)) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, funcPart(O))$

Task-based Social Rules:

$r^{9}_{[0.8,0.9]} := near(O, G, holdPart(O)) \wedge$

$task(X, use, O) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, holdPart(O))$

$r^{10}_{[0.8,0.9]} := near(O, G, funcPart(O)) \wedge$

$task(X, give, O) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, funcPart(O))$

$r^{11}_{[0.95,0.95]} := near(O, G, holdPart(O)) \wedge$

$\neg dirty(O)$

$task(X, use, O) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, holdPart(O))$

$r^{12}_{[0.95,0.95]} := near(O, G, funcPart(O)) \wedge$

$\neg dirty(O)$

$task(X, give, O) \wedge$

$domain(X, kitchen) \implies$

$graspable(X, O, funcPart(O))$

Object Interaction Rules:

$r^{13}_{[0.8,0.9]} := grasped(O, holdPart(O)) \wedge$

$task(X, use, O) \wedge$

$domain(X, kitchen) \implies$

$useable(X, O)$

$r^{14}_{[0.8,0.9]} := grasped(O, funcPart(O)) \wedge$

$task(X, give, O) \wedge$

$domain(X, kitchen) \implies$

$giveable(X, O, H)$

$r^{15}_{[0.55,0.95]} := grasped(O, holdPart(O)) \wedge$

$domain(X, kitchen) \implies$

$setOnTable(X, O, T)$

$r^{15}_{[0.55,0.95]} := grasped(O, funcPart(O)) \wedge$

$domain(X, kitchen) \implies$

$setOnTable(X, O, U)$

$r^{17}_{[0.8,0.9]} := grasped(O, holdPart(O)) \wedge$

$dirty(O) \wedge$

$domain(X, kitchen) \implies$

$setOnTable(X, O, U)$

$r^{18}_{[0.8,0.9]} := grasped(O, funcPart(O)) \wedge$

$dirty(O) \wedge$

$domain(X, kitchen) \implies$

$setOnTable(X, O, U)$

Rules $r^1 - r^3$ relate to the agent's general commonsense understanding of the physical properties of objects. E.g., sharp edges provide a cutting affordance.

Rules $r^4 - r^6$ prescribe several social rules in a kitchen environment. For example, dirty objects need to be sanitized and only objects not currently used by someone else are available for grasping. We note that there may make be exceptions to these rules, for example, grasping the object (like a mixer) at a functional part may still be infeasible or inappropriate. Moreover, even if the functional part is graspable, extra care may need to be exercised when doing so.

Rules r^7 and r^8 provide general rules on grasping objects. For example, if there are grasp points near the handle of a knife, then the handle has a grasp affordance, and if there are no grasp points near the handle, but there are some near the blade, then the blade has a grasp affordance.

Rules $r^9 - r^{12}$ relate to social etiquette and convention when using and giving objects. These rules provide a narrowing context depending on the agent's current task of using the object itself or giving the object to another. For example, when handing over an object, it is "proper" to hold the functional part of the object (e.g., knife) and present the handle towards the recipient.

Rules $r^{13} - r^{18}$ are kitchen rules that apply when the agent is manipulating and interacting with the object. For example, if the agent is holding the knife by its

blade and is tasked with giving it to a human, then it is afforded the possibility of giving the knife.

4.4.6 Handover Inference – Introduction

We will now turn to how a robot with our computational framework can use these rules to reason through the process of handing over a knife. Consider a robot helper receiving instructions from a human, Julia. Suppose Julia says to the robot: *"Bring me something clean I can use to cut this tomato."* The robot will need to parse this request and infer affordances of objects in its environment in context. Before we can describe how the robot can perform this inference for the knife-handover example, we will first describe our inference process and algorithm more generally.

4.4.7 Inferring Affordances with Uncertain Logic

The goal of defining cognitive affordance models is to infer object affordances based on (1) their perceivable features, (2) the known context, and (3) general domain and common sense knowledge. We propose to start with the first prototype inference algorithm shown in Algorithm 6.1 and refine it to tailor it specifically to a cognitive affordance model. The algorithm takes three parameters: (1) a BoE of candidate perceptions $\{\Theta_F, m_f\}$ is provided by the low-level vision system, (2) a BoE of relevant contextual items $\{\Theta_C, m_c\}$ provided by a knowledge base or some other part of the integrated system that can provide context information, and (3) a table of cognitive affordance rules R. Each rule $r_{f \wedge c \to a}$ in R is indexed by a feature perception f and a set of contextual items c, and dictates the mass assigned to $Bel(a)$ and $Pl(a)$ when the system believes the degree to which object features f were detected and that contextual items c are true. Here, a is a complex logical expression representing the affordance that can be derived from the perceived features f in context c.

The inference algorithm then examines each rule $r_{fc \to a} \in R$ (line 5), and m_{fc} is determined by performing $m_f \otimes m_c$ (line 6), where m_f specifies the degree to which object feature f is believed to be detected, and m_c specifies the degree to which each of the rule's associated contextual items is believed to be true. *Uncertain*

Algorithm 4.1 getAffordance($\{\Theta_F, m_f\}$,$\{\Theta_C, m_c\}$,R)

1: $\{\Theta_F, m_f\}$: BoE of candidate perceptual features
2: $\{\Theta_C, m_c\}$: BoE of relevant contextual items
3: R: Currently applicable rules
4: $S = \emptyset$
5: **for all** $r \in R$ **do**
6: $\quad S = S \cup \{(m_f \otimes m_c) \odot m_{r=fc \to a}\}$
7: **end for**
8: $G = group(S)$
9: $\psi = \emptyset$
10: **for all** group $g_a \in G$ **do**
11: $\quad \psi = \psi \cup \{\bigcap_{j=0}^{|g_a|} g_{a_j}\}$
12: **end for**
13: **return** ψ

Modus Ponens is then used to obtain m_a from $m_{fc \to a}$ and m_{fc} (line 6).

Note that since we allow multiple affordance rules to be considered, multiple affordances may be produced. Multiple rules may produce the same affordances for various reasons, possibly at different levels of belief or disbelief. However, we seek to return the set of *unique* affordances implied by a set of perceptions f.

After considering all applicable affordance rules, we group affordances that have the same content but different mass assignments (line 8), and use Yager's rule of combination (line 11) to fuse each group of identical affordances, adding the resulting fused affordance to set ψ. This set then represents the set of affordance implied by the perceived features f.

Finally, we can use the confidence measure λ to determine whether an inferred affordance should be realized and acted upon. For example, we could check the confidence of each affordance $a \in \psi$ on its uncertainty interval $[\alpha_i, \beta_i]$: if $\lambda(\alpha_i, \beta_i) \leq \Lambda(c)$ (where $\Lambda(c)$ is a confidence threshold, possibly depending on context c), we do not have enough information to confidently accept the set of inferred affordances and can thus not confidently use the affordances to guide action. However, even in this case, it might be possible to pass on the most likely candidates to other cognitive systems. Conversely, if $\lambda(\alpha_i, \beta_i) > \Lambda(c)$, then we take the inferred affordance to be certain enough to use it for further processing.

4.4.8 DS-Theoretic Handover Inference

Returning to our knife-handover example, the robot, $X = self$, parses the request from Julia (i.e., for an object she can use to cut a tomato) and assigns its own task context and determines the types of affordances it is interested in exploiting in the kitchen environment. The agent is confident that it is in the kitchen context and that it is in the context of handing over an object in the kitchen, and assigns context masses as follows:

$domain(self, kitchen)$: $m_{c_{1,1}} = 1.0$
$task(self, give, O)$: $m_{c_{2,1}} = 0.95$
$task(self, use, O)$: $m_{c_{2,1}} = 0.05$

Specifically, the robot's cognitive architecture includes a natural language processing component, which processes Julia's instruction. The phrase "bring me something" is taken by the robot to indicate a "give" context as opposed to a "use" context. This contextual information is obtained outside of CALyX and passed into it as input. Similarly, the robot's cognitive architecture includes belief and goal management components, which process the robot's current role as a kitchen helper and compute the likelihood that is in the "kitchen" domain. This, too, is passed as input to our CALyX system.

CALyX's Memory Management subcomponent receives this contextual information and selects applicable rules (Fig. 4.3, step 1). LTM potentially contains a large set of rules across various contexts that the robot has acquired over its lifetime. Given the specific domain and task contexts, the Memory Management subcomponent selects a subset of applicable rules from the LTM (step 2) and populates the WM with these rules (step 3).

Although the context has been established and the applicable rules have been identified, at this point the robot is not yet ready to do any affordance inference because it does not know whether the perceptual aspects specified by the rules are satisfied. Given the set of applicable rules in WM (Fig. 4.4, step 1), PAC can guide or direct sensory processing systems (step 2) to determine the uncertainties

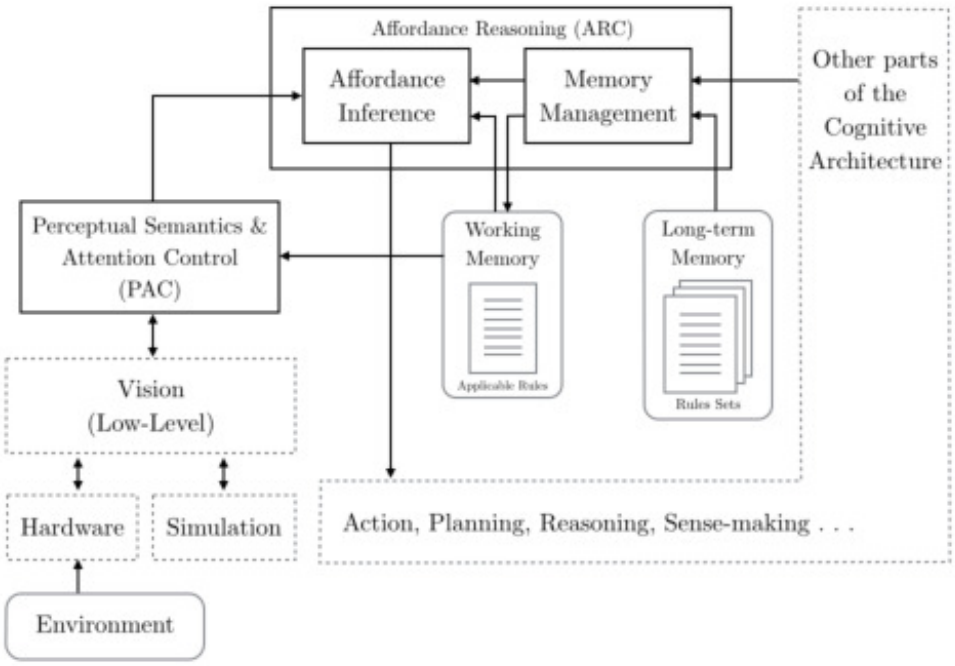

Figure 4.2: Computational Architecture (CALyX). We depict our contribution in bold solid lines.

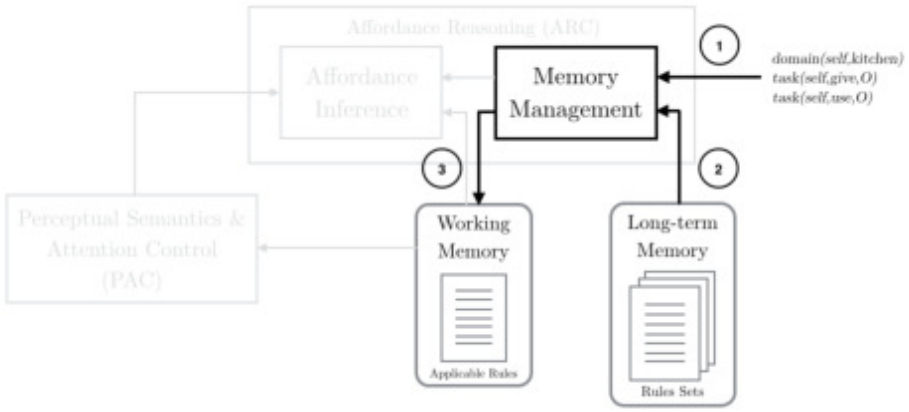

Figure 4.3: Selecting Applicable Rules Based on Context

associated with each of the perceptual aspects specified in the set of applicable rules (step 3). For example, based on its set of rules, it knows to look specifically for certain visual percepts relevant to the rules, such as *near*(), *sharpEdge*(), and so on. Note, at this point, the robot is not aware of a specific object that it needs to find, but with PAC in combination with the low-level vision system, it can scan its environment and examine each object more closely to determine which perceptual aspects specified above are satisfied.

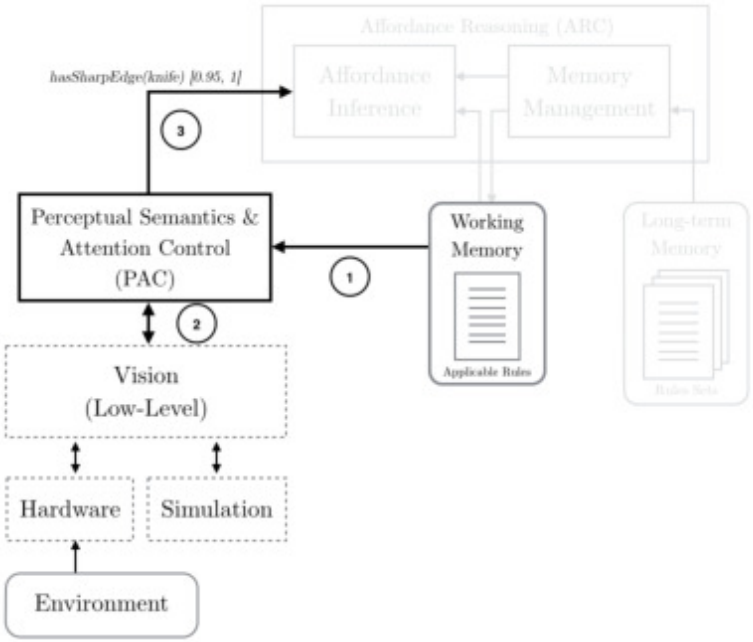

Figure 4.4: Determining Beliefs of Perceptual Aspects

4.4.8.1 Spot the Knife - Directed Perception with PAC

Let us assume that the agent spots a knife on the counter. Upon examining the physical features of the knife, the agent determines masses for percepts in each of its perceptual aspects.

For example, upon spotting the knife, PAC assigns the percept $hasSharpEdge(knife)$ with a mass $m_{f_{3,1}} = 0.95$. However, because it is slightly unsure it also assigns the possibility that the knife either has or does not have a sharp edge, $\{hasSharpEdge(knife), \neg hasSharpEdge(kn$ with a mass = 0.05. With these two masses, support for the percept $hasSharpEdge(knife)$

falls within the interval $[\alpha, \beta] = [0.95, 1]$. Similarly, we compute uncertainty intervals for all the relevant percepts, when the agent sees the knife:

$hasSharpEdge(knife)$ [0.95, 1]

$hasPointyTip(knife)$ [0.8, 0.9]

$hasOpening(knife)$ [0, 0]

$near(knife, G, holdPart(knife) = handle)$ [0.95, 0.95]

$near(knife, G, funcPart(knife) = blade)$ [0.95, 0.95]

$grasped(knife, holdPart(knife) = handle)$ [0, 0]

$grasped(knife, funcPart(knife) = blade)$ [0, 0]

$dirty(knife)$ [0.31, 0.81]

$inUse(knife, H)$ [0, 0]

To summarize, the agent has detected a knife that has a sharp edge and can be grasped, but it is not entirely sure if it is clean or dirty. Note, the agent has yet to pick up and grasp the object, so the *grasped*() predicates evaluate to $[0, 0]$, which means *logically false* with maximum certainty.

4.4.8.2 Affordance Inference with ARC

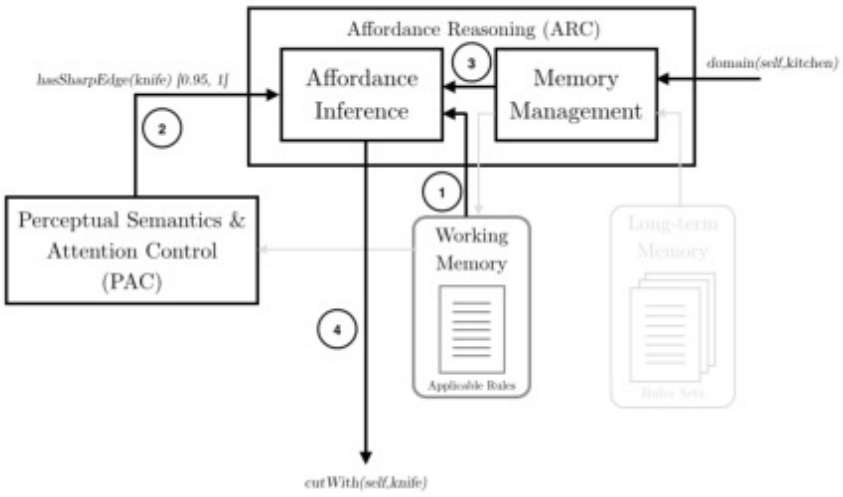

Figure 4.5: Performing Inference of Cognitive Affordances

ARC examines each rule $r^i_{[\alpha_i, \beta_i]}$ (line 5 of Algorithm 1) in WM (Fig. 4.5, step

1), and m_{fc} is determined by performing $m_f \otimes m_c$ (line 6), where m_f specifies the degree to which the percept f is believed to be observed (Fig. 4.5, step 2), and m_c specifies the degree to which each of the rule's associated contextual value is believed to be true (Fig. 4.5, step 3). DS-based modus ponens is then used to obtain m_a from $m_{fc \to a}$ and m_{fc} (line 6) (Fig. 4.5, step 4).

For example, consider rule r^1:

$r^1_{[0.8,1]} := hasSharpEdge(O) \wedge domain(X, kitchen) \implies$
$cutWith(X, O)$

The agent will apply perceptual and contextual information as follows, to determine the affordance implied by the rule:

$r^1_{[0.8,1]}(m_r = 0.8) :=$
$hasSharpEdge(knife)(m_f = 0.95) \wedge$
$domain(self, kitchen)(m_c = 1.0) \implies$

$cutWith(self, knife)(m_a = (m_f \otimes m_c) \odot m_r = 0.76)$

The uncertainty interval for the rule can then be computed as $[0.76, 1]$.The agent will subsequently perform this analysis for each of the other rules in the set to determine uncertainty intervals for the implied affordances.

To be able to generate a set of unique affordances, a, implied by feature, f, after considering all applicable affordance rules, we thus group affordances that have the same semantic content but different mass assignments (line 8) and use Yager's rule of combination (line 11) to fuse each group of identical intentions, adding the resulting fused intention to set ψ. Thus, affordances from rules r^1, r^3, r^6 and r^{17} will be fused as these rules all apply to the affordance of $graspable(self, knife, holdPart(knife) = handle)$. The agent will also fuse together rules r^2, r^4, r^7 and r^{18} as these rules all apply to the affordance of $graspable(self, knife, funcPart(knife) = blade)$. The agent will further fuse together rules r^{13}, r^{14}, r^{15} and r^{16} as these rules all apply to the same affordance of $setOnTable(self, knife, table)$.

Based on the application of each rule to the semantic visual percepts and contextual items, and fusing rules with similar implied affordances together we can generate a list of unique affordances available to the agent at the current moment in time, when it has seen the knife:

Available affordances (Upon seeing the knife), ψ
$cutWith(knife)[0.76, 1], \lambda = 0.29$
$pierceWith(knife)[0.64, 1], \lambda = 0.16$
$containWith(knife)[0, 1], \lambda = 0$
$sanitizeable(knife)[0, 0.95], \lambda = 0.004$
$graspable(self, knife,$
$holdPart(knife) = handle)[0.96, 0.99], \lambda = 0.78$
$graspable(self, knife,$
$funcPart(knife) = blade)[0.98, 0.99], \lambda = 0.88$

This information is passed from CALyX to other parts of the cognitive architecture like the robot's goal and action management system, which performs different operations based on these measured uncertainty intervals and associated λ confidence measures. The agent might decide that because there is a high degree of confidence that the object under consideration has a $cutWith$ affordance, it will choose to grasp it and then select to grasp it at the blade, given the context of a handover. Because it is unclear that the knife is dirty, the agent is less confident that the knife needs cleaning.

4.4.8.3 Grasp the Knife - Iterated Directed Perception with PAC

Affordance inference in this manner with CALyX is an iterative or cyclical process and affordances are computed and re-computed continuously to guide the robot's next actions. Thus, once the robot has grasped the knife, CALyX is once again tasked with inferring affordances to determine what the robot can and cannot do with the knife that it is holding. Here the contexts remain the same, so the rules in WM are likely unchanged. However, because the world state has changed (i.e.,

the knife is no longer on the table but in the hands of the robot), the perceptual aspects specified by the rules potentially have different truth values. PAC is once again called upon to update the uncertainty intervals associated with the visual perceptual aspects as follows:

$hasSharpEdge(knife)$ [1, 1]

$hasPointyTip(knife)$ [0.95, 0.95]

$hasOpening(knife)$ [0, 0]

$near(knife, G, holdPart(knife) = handle)$ [0.95, 0.95]

$near(knife, G, funcPart(knife) = blade)$ [0.05, 0.05]

$grasped(knife, holdPart(knife) = handle)$ [0, 0]

$grasped(knife, funcPart(knife) = blade)$ [1, 1]

$dirty(knife)$ [0.95, 1]

$inUse(knife, self)$ [1, 1]

It updates the $grasped()$ predicate information because it has now grasped the knife's blade. The robot also detects that the knife is dirtier than initially determined, and there are no longer any grasp points available near the blade, because it is already holding the blade. It also is more certain that the knife has a sharp edge and pointed tip. Finally, it also knows that because it is using the knife, the knife is "inUse".

4.4.8.4 Iterated Affordance Inference with ARC

Based on this update, ARC will re-calculate the uncertainty intervals and confidence measures associated with its affordances, as follows:

Available affordances (after grasping knife), ψ
$cutWith(knife)[0.8, 1], \lambda = 0.34$
$pierceWith(knife)[0.76, 1], \lambda = 0.29$
$containWith(knife)[0, 1], \lambda = 0$
$sanitizeable(knife)[0.9, 1], \lambda = 0.564$
$graspable(self, knife,$ $holdPart(knife) = handle)[0.5, 0.9], \lambda = 0.07$
$graspable(self, knife,$ $funcPart(knife) = blade)[0.03, 0.9], \lambda = 0.0006$
$setOnTable(self, knife, table)[0.84, 0.99], \lambda = 0.43$
$useable(self, knife)[0, 0.95], \lambda = 0.004$
$giveable(self, knife, Julia)[0.9, 1], \lambda = 0.56$

Having detected that the knife is dirtier than initially determined, the agent now has a higher confidence that the knife has a sanitizeable affordance. The agent also has additional affordances available. It has high confidence that the knife is giveable to Julia and that the knife can be set on the table. It knows that the knife is not currently useable to cut things by itself, mainly because it is holding the blade and the current context is a handover, and not use.

Once again, this information is passed to other parts of the agent's cognitive architecture like goal and action management systems, which perform different operations based on these measured uncertainties. The agent might decide to choose to realize one or more of the above observed affordances.

4.5 Experiment - Multi-Domain, Multi-Scenario Handover

4.5.1 Introduction

In our first example of a kitchen helper handing over a knife, we demonstrated the capability of our framework to reason about cognitive affordances of handing over objects. We limited the experiment to one domain (i.e., the kitchen helper) and

we focused on a simple set of rules that govern social interactions in this domain. We demonstrated a flexible reasoning process that took into account social context. In this second experiment, we extended the object handover task and compared interactions across various domains and expanded our notion of object affordances to "social affordances" offered by humans in the scenario, as well. Object handovers are often complex interactions that involve more social intelligence than just reasoning about physical or social aspects of the objects alone. Often, in human-human interaction scenarios, the giver must tune into various social cues (e.g., eye-gaze) offered by the receiver indicating whether a handover must be initiated or not. Social context is highly relevant to object handovers, and we will demonstrate the flexibility of our framework in reasoning about situations when similar observations of the environment have very different meanings in different contexts.

For this experiment we built on some of the extensive previous work in determining parameters of a handover from a physical and temporal sense and in deciding its timing and trajectories [GHB+10, STS95]. With regards to social cognition during handovers, Strabala et al. have extensively studied social cues that are crucial to coordinating a handover [SLD+13]. They provide four exemplary domains – elder care-giver, mechanic-helper, fire-brigade volunteer, and flyer-handout giver – that all feature a handover activity, but under very different contexts. In this experiment, we implemented an affordance-based handover reasoning mechanism for these four socially distinct domains. While Strabala et al. focused on discovering a unified handover structure that might apply in all these four domains, we retained the richness and distinctions of these four different domains, and instead reasoned about the affordance of "transferability" of an object prior to the handover. We will begin by describing the four domains in more detail.

4.5.2 Domains

We will consider four exemplary domains as originally introduced by Strabala et al., and expanded by us, as follows:.

1. **Care-giver** at an elder care facility: In this domain, a care-giver or assistant holding a glass of water is tasked with handing it over to the patient. The care-giver must be sure that the patient is ready to receive the water before beginning the transfer process. In many cases, it is further desirable that the patient make eye contact and orient her body towards the care-giver. Moreover, in this scenario, it is often not appropriate for the care-giver to handover the water when the patient is not attentive and looking away, even if the patient is reaching out or verbally requesting the water.

2. **Mechanic's Helper:** In this domain, a helper is tasked with handing over a wrench to a mechanic. The social cues in this domain, while similar in structure to the cues in the care-giver scenario, are vastly different in how they influence the interaction. Here, the mechanic may be under a car or focused on the task at hand, and, therefore, not attentive to the assistant. So the helper must be more attentive to other signals such as the mechanic reaching out with an outstretched arm and verbally requesting the wrench. Eye contact and body orientation may be less important in this scenario. Indeed, sometimes the mechanic may be facing the helper but not be ready for the wrench, hence the handover must be confirmed through verbal signals or by reaching out.

3. **Fire-brigade:** In this domain, a helper or volunteer, who is one of many agents (humans and robots) in a line, is tasked with passing buckets of water from a source to the scene of a fire. This domain is different from the prior two domains in its ignorance of social context. Generally, there is a known procedure for swinging buckets and the urgent nature of the situation has eliminated the role of social etiquette. In this domain, it is often permissible to begin a handover procedure without many social cues like eye gaze, reaching, or verbalizing. Bodies are often not facing each other and the only real condition to begin the transfer is possession of the bucket of water. As long as the giver is holding a bucket of water, the transfer should begin.

4. **Flyer-handouts:** In this domain, a giver is tasked with handing out flyers on a busy university sidewalk. As Strabala et al. note, the giver has no prior relationship with the people on the sidewalk, and so established social norms apply [SLD+13]. In fact, this domain is the complement of the Fire-brigade domain because many social cues, including eye gaze, body orientation, reaching out actions and verbal confirmation, all apply. The passerby who is interested in receiving a flyer is likely to face the giver, make eye contact and request a handout while reaching out. The giver would be considered rude if she imposed a flyer on someone who was merely walking towards her or if the passerby provided verbal cues and hand motions that might appear to be a reaching action, when in fact they were signaling the opposite.

4.5.3 Representing Social Cues and Domain Rules

To represent these domains, we first selected well-established social cues offered by the receiver to the giver that are often considered relevant to handovers: eye gaze or eye contact, verbal confirmation, the action of reaching out and requesting the object, and body orientation [SLD+13]. We represented these social cues as predicates (with intuitive semantics) $eyeGaze(X)$, $verbalSignal(X)$, $reachingOut(X)$, and $bodyFacing(X)$, respectively, where X refers to the receiver. In addition to the social cues, we also represented the information that the robot is holding object O (i.e., the object to be handed over) with the predicate $holding(self, O)$. We represented the affordance of transferability with the predicate $transferable(O, X)$ to mean that object O is transferable to receiver X.

With these social cues, we assigned simple social rules or norms that are applicable generally (but to varying degrees) across all four domains, as follows:

$$r^1 := eyeGaze(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$$

$$r^2 := verbalSignal(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$$

$r^3 := reachingOut(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$

$r^4 := bodyFacing(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$

$r^5 := holding(self, O) \wedge goal(handover) \implies transferable(O, X)$

$r^6 := eyeGaze(X) \wedge bodyFacing(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$

$r^7 := verbalSignal(X) \wedge reachingOut(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$

$r^8 := eyeGaze(X) \wedge bodyFacing(X) \wedge verbalSignal(X) \wedge reachingOut(X) \wedge holding(self, O) \wedge goal(handover) \implies transferable(O, X)$

4.5.4 Representing the Domain Distinctions

We represented the different levels of importance of the rules in the four domains by assigning them different uncertainties depending on the domain. We selected uncertainties for the rules from three settings, as follows:

High = [0.95, 1]

Medium = [0.5, 0.6]

Low = [0.31, 0.81]

Here, “High" refers to the setting in which the rule is believed to be true with a low uncertainty (i.e., high degree of certainty). Similarly, the “Medium" and “Low" levels each correspond to a setting in which the rules are believed to be true, but with medium and high uncertainty, respectively.

We assigned an uncertainty level for each of the rules in each of the domains, as shown in Table 4.1. For example, since all social cues are important for the Flyer-handout domain, rule 8 featuring all of these cues is the most important one. In

contrast, in the Fire-brigade domain, the only rule that is the most important is the commonsensical rule 5, which requires possession of the bucket before transfer, without any other social overhead. In the Elder care domain it is important that the patient face the giver and make eye contact, whereas in the mechanic domain, it is important that the mechanic reach out and ask for a wrench before initiating handover. Accordingly, rules corresponding to these social cues were given more importance and higher certainty.

4.5.5 Experimental Scenarios

For our experiment, we considered four scenarios that are possible within any or all of these domains. These four scenarios represent the truth settings for the set of perceptual social cues - eye gaze, body orientation, verbal request or reaching out - in a given situation. For example, a mechanic under a car requesting a wrench may not make eye contact or orient their body towards the giver, so the perceptual cues like eye gaze and body orientation would be False in this scenario. But the mechanic is reaching out his arm and requesting a wrench verbally, so the perceptual cue for verbal request and reaching out would be True. We acknowledge the fact that with four social cues, there are 16 possible scenarios. In particular, each scenario would include a combination of true/false settings for the four perceptual cues: eye gaze, body orientation, verbal request or reaching out. However, we have limited our

Rule	**Elder Care**	**Mechanic**	**Fire-Brigade**	**Flyer-Handout**
r^1	Medium	Low	Low	Low
r^2	Low	Medium	Low	Low
r^3	Low	Medium	Low	Low
r^4	Medium	Low	Low	Low
r^5	Low	Low	High	Low
r^6	High	Low	Low	Medium
r^7	Low	High	Low	Medium
r^8	Low	Low	Low	High

Table 4.1: Domain-Specific Rule Uncertainty Assignment

Scenarios	Eye Gaze	Verbal Signal	Reaching Out	Body Facing	Holding
S^1	T	F	F	T	T
S^2	F	T	T	F	T
S^3	T	T	T	T	T
S^4	F	T	F	F	T
$S^1(uncertain)$	sT	sF	sF	sT	sT
$S^2(uncertain)$	sF	sT	sT	sF	sT
$S^3(uncertain)$	sT	sT	sT	sT	sT
$S^4(uncertain)$	sF	sT	sF	sF	sT

Table 4.2: Perceptual Uncertainty Assignment for 4 Scenarios

presentation in this chapter to the four following exemplary scenarios that were the most informative, across all our domains:

Scenario 1 (S^1): When the receiver is making eye contact and facing the giver, but not providing any verbal cues or reaching out to the giver (e.g., an elderly patient signaling that they are ready to receive water).

Scenario 2: (S^2): When the receiver is verbally requesting the object from the giver and is extending her arm in anticipation of receiving the object. The receiver, however, is busy performing another task and is turned away and not looking at the giver.

Scenario 3: (S^3): Here the receiver is fully engaged in the handover and is providing all social cues.

Scenario 4: (S^4): Here, the receiver is only signaling verbally that he is ready to receive the object, but his attention maybe elsewhere as he is looking and turned away from the giver. He is also not extending his arm or reaching out to the giver.

We also generated four additional scenarios that were variations of the first four. These additional scenarios only differed from the original four in that we diminished the truth and false certainties using the "Somewhat True" and "Somewhat False" uncertainty setting:

True (T) = [0.95, 1]

False (F) = [0, 0.05]

Somewhat True (sT)= [0.62, 0.96]

Somewhat False (sF) = [0.04 0.38]

Here, the “Somewhat True" is an uncertainty setting for perceptual cues that are logically true, but with a greater amount of uncertainty than that of the “True" setting (i.e., wider uncertainty intervals). Similarly, the “Somewhat False" is an uncertainty setting for perceptual cues that are logically false, but with a greater amount of uncertainty than that of the “False" setting.

Overall, we represented the truth of the social cues in all scenarios per uncertainty levels noted above and shown in Table 4.2. Note, these scenarios represent a truth setting for each of the perceptual cues in a situation. The scenarios themselves are independent of the domain. Thus, for example, in a domain, the giver could perceive a set of cues whose truth settings could fit any one of the scenarios.

4.5.6 Experimental Results and Discussion

We performed affordance inferences for a total of 32 situations involving our eight scenarios (four certain and four uncertain) across four domains. The results shown in Fig. 4.6 depict the confidence measure across the various scenarios. Note, that higher λ values indicate a tighter uncertainty interval and consequently a more confident or clear outcome. For each scenario, we have depicted blue and yellow plots corresponding to whether or not the uncertainty setting was True/False or Somewhat True/Somewhat False, respectively. For example, in Scenario 1, the blue plots correspond to the case when the $eyeGaze()$ and $bodyFacing()$ cues were assigned a setting of True = [0.95, 1]. In this scenario, the yellow plots correspond to the case when the $eyeGaze()$ and $bodyFacing()$ cues were assigned an uncertainty setting of Somewhat True = [0.62, 0.96].

The results show that our computational framework conforms to our intuitions about these various scenarios. Specifically, we show some interesting distinctions between various domains. For example, in Scenario 2 (S^2), a transferable

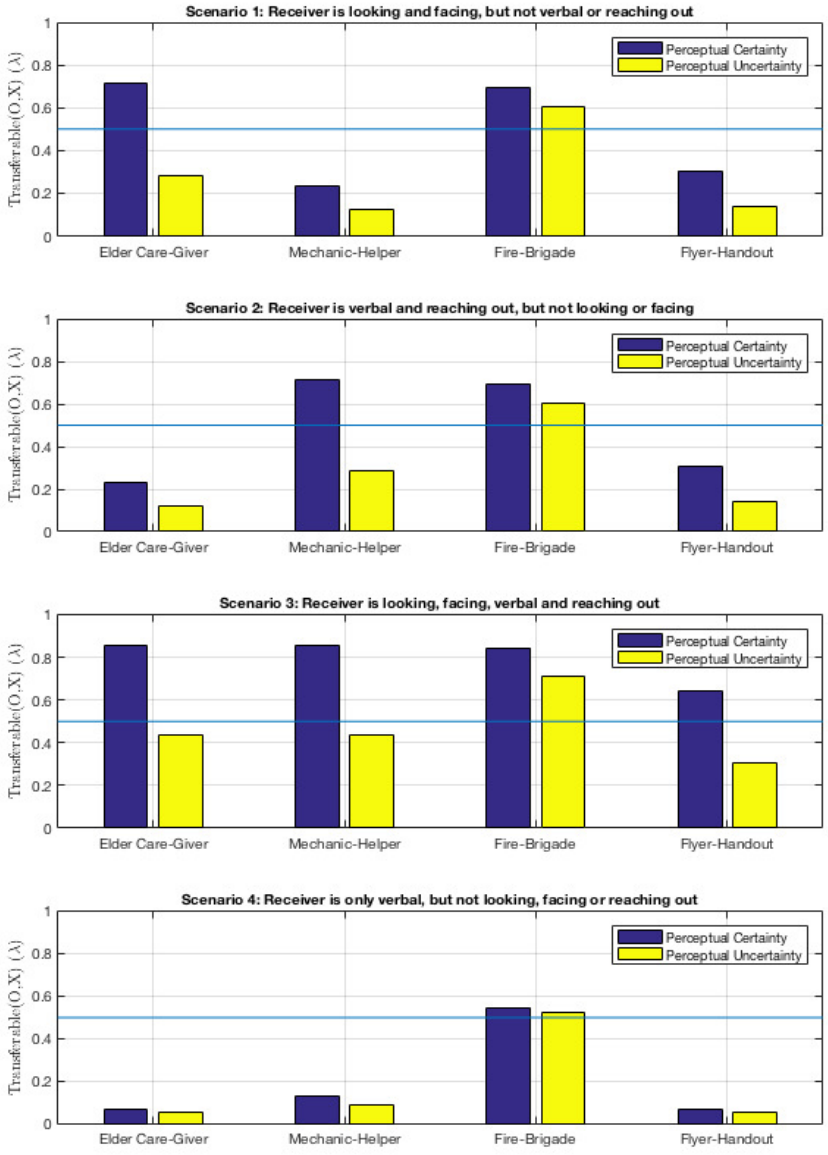

Figure 4.6: Plots showing whether an object being held by the robot-giver has an affordance of being "transferable" to the receiver. Each plot corresponds to a particular perceptual scenario (truth values for the existence of various social cues) across four domains - Elder Care-Giver, Mechanic-Helper, Fire-Brigade Volunteer and an agent handing out Flyers. The perceptual cues in a scenario have assigned truth values and an uncertainty interval. The yellow plots correspond to the case when the cues have a higher level of uncertainty. The experiment was performed for four different scenarios and the results correspond to expected human-human interaction behavior in these scenarios and domains.

affordance is considered to exist when a mechanic extends his arm and requests a wrench even if he is not looking at or facing the robot. While, that same type of behavior from a patient in an elder-care facility may not offer the same kind of transferability affordance. We further note that the transferability affordance is always present across various scenarios in the Fire-brigade domain, while it is only present for the Flyer-handout domain when the receiver is fully engaged with the giver. As noted earlier the Fire-Brigade domain represents a domain without much social context and so it is expected that regardless of the social cues, buckets must be passed along. On the other hand, the Flyer-Handout domain represents a domain with deep social context and so it is expected that a transfer should not be initiated unless all social cues are present.

We also show what happens when there is uncertainty introduced to the various perceptions (see yellow plots in Fig. 4.6). It may be uncertain as to whether the receiver is facing the robot or has verbally requested the object. There may also be uncertainties associated with the amount of eye contact or whether an extended arm actually means that the receiver is "reaching out". Interestingly, even with uncertainty the overall structure of the results remained the same but the confidence of the conclusions dropped. That is, the robot is less sure of whether an object has a transferability affordance when it was less certain about whether or not it was perceiving the social cues correctly. In these situations, we can have the robot clarify the social cues and hold back and wait to handover the object. Note, however, the Fire-brigade domain exhibited the least change even in the face of uncertainty. This is because we emphasized the limited role played by social cues in these emergent and largely procedural handover domains.

4.6 Discussion

4.6.1 Implications of the Proposed Architecture

The novel CALyX architecture for inferring cognitive affordances has an affordance-based reasoning component (ARC) and an attention control and perceptual seman-

tics component (PAC). In much of the past work, affordances have been treated as a subsystem of vision or planning. CALyX allows affordance processing to be a separate component, not subsumed in a vision system or a planning engine. The robot's vision system is then just one of several sensory systems that work with affordance processing. For example, certain perceptual aspects like an object's weight may need the robot's haptic and touch systems to resolve. Thus, an affordance rule involving weight might not involve the vision system at all, and consequently PAC will need to direct the robot's grippers and arms to determine weight information.

In CALyX, affordance rules inform and guide visual attention, that is PAC is guided by the perceptual aspects present in WM at any given moment in time. There have been findings in psychology to support this approach and these findings suggest that affordances influence visual attention by biasing and focusing the visual search on those objects that afford relevant actions. Particularly, researchers have shown support for top-down control in attention processing, task-based priming of visual search [Yar67, NI05, PRZV14], and the influence of affordances on visual attention [TM14, RH11, GVS14]. Developmentally, this is very advantageous as well because as the robot develops, our architecture allows the robot to take into consideration additional perceptual aspects and social cues as it learns them. These social cues will present themselves in the rules and allows the robot to reason about these cues in relation to others as we have shown here. Moreover, the framework allows the robot to revise its rules and beliefs without the need to undergo new rounds of training and learning, as is typically needed for other statistically-inclined affordance learning frameworks.

We have also shown ARC and PAC as separate components of CALyX. ARC and PAC perform different functions, as noted above. However, this separation is not merely a functional one. Each component is autonomous and does not rely on the other beyond an input-output relationship. PAC can potentially interface with not only the vision system but also a haptic system and auditory system to perform perceptual semantic analysis and attention on these other modalities when the affordance rules demand it. Similarly, ARC interfaces with various other higher-level

cognitive systems (planning and reasoning) to perform affordance-based reasoning tasks.

CALyX is a flexible architecture and does not preclude maintaining an episodic memory of situations that involve acting on certain affordances and observing the effects. For example, although not described explicitly, CALyX does not exclude the robot from tracking and maintaining the effects of grasping the knife and handing it over to the human. We noted earlier, that the truth values of the perceptual aspects can change from moment to moment and CALyX does not preclude tracking this information. In fact, observing the effects can influence the uncertainty of some rules in the current context. In the future, when the robot encounters similar contexts, it can remember these rules and reason accordingly.

We recognize that because our framework is rule-based, there is a possibility for rule conflicts. Rule conflicts can arise in many ways. One way is if a feature and its negation produce the same affordance. If there is a rule ($r_1 :\equiv handle \wedge context_1 \implies graspable$), then there might be a conflict if there is then another rule that states that ($r_2 :\equiv \neg handle \wedge context_1 \implies graspable$). Conflicts of the type between r_1 and r_2 are currently being resolved in our framework through our Yager fusion operator (which was designed for handling conflicting evidence of this sort) combining uncertainty evidence for *graspable* from each of rules r_1 and r_2. Another way a conflict may arise is if the same feature in the same context can produce conflicting affordances. For example, if in addition to rule r_1, there is another rule implying a negation of an affordance, like ($r_3 :\equiv handle \wedge context_1 \implies \neg graspable$). Currently, in the examples we have presented, we have not explicitly reasoned about "negative affordances" like $\neg graspable$. However, it is reasonable to expect that both *graspable* and $\neg graspable$ could belong to the same frame of discernment, and consequently, fusing conflicting evidence from rules r_1 and r_3 can be handled in a similar manner.

Overall, a main advantage of the proposed architecture is that rational choices can still be made in the face of conflict because of the underlying character of the uncertain logic based inference algorithm. That is, the algorithm considers the rules (conflicting or otherwise) together through the fusion operator and collectively deter-

mines implications. Moreover, the architecture could be coupled with a higher-level predictive engine to test expectations against observations and adjust rule uncertainties.

Our architecture does not preclude, and in fact encourages the selection of an optimal choice of affordance, especially when there are many choices available. This is because at any given moment, the architecture presents a set of affordances along with uncertainty intervals. It makes no judgment about which one to select, because that is not the function of the affordance inference process. Selecting an optimal affordance to act on is the job of other cognitive functions like planning and goal management. CALyX provides helpful metrics such as the λ confidence measure, but it does not specify any further requirement in regards to selecting certain affordances.

One concern is that as the number of rules increase, there is a potential for higher time and space complexity. The architecture does not explicitly address this beyond suggesting the use of a limited working memory to track applicable rules and only perform inference on this reduced rule-set.

CALyX is capable of withstanding changes in the environment and can adjust for these changes over its cognitive cycles. Short term changes in the environment can impact the agent's decision process. The architecture does not preclude adapting to current demands, even if that means pursuing short term changes temporarily. This is because, the architecture is more focused on moment-to-moment or cycle-to-cycle affordance inference and not more general planning and goal management issues.

Note that the proposed architecture does not preclude the agent from operating with a certain degree of autonomy. With the inclusion of an uncertainty interval and the confidence measure, not only does the architecture allow the agent to ask clarification questions when certain aspects are unclear, but we can also track and quantify the level of autonomy for the agent based on how frequently it encounters ambiguity. An agent with one too many uncertain rules or an agent with faulty sensors that result in uncertain percepts is less autonomous. This allows for the agent's

general reasoning abilities to be quantifiable.

We should note that we have implemented our algorithm and the CALyX architecture in connection with a robotic vision system, and we have integrated it into the larger DIARC architecture [SSKA07], although these additional aspects are beyond the scope of this chapter.

4.6.2 Learning the Rules

Thus far, we have not discussed, explicitly, the origin of the cognitive affordance rules and how an agent might generate or learn new rules, because this is not the focus of this chapter. Our focus in this chapter, instead, was to demonstrate our affordance representation format, inference algorithm, and architecture. Nevertheless, we expect these rules can be learned in a number of different ways from observation, demonstration and exploration, and using multiple different modalities including vision, natural language and haptic information. The agent could learn these types of rules from explicit natural language teaching and instruction as shown by Cantrell et al. [CTS+12]. The agent could also learn various rules from observation through reinforcement learning (RL) methods as shown by Bouralias [BBS15] or through exploration from methods as shown by Forestier [FP15]. Alternatively, the agent could also acquire these rules through data mining and various association rule-mining techniques [WKPW09]. Techniques for learning rules, especially those associated with social norms, is the subject of Part II of this dissertation.

4.7 Using Inferred Affordances - Future Work

The focus of the proposed architecture is affordance inference (what to do with the inferred affordances is beyond the scope of this chapter). Generally, the architecture leaves open the possibility for how the affordance information can be utilized suitably. As discussed before, these affordance computations are useful in planning problems and in guiding a robot's next actions. In fact, the benefits of affordance computations extend further and as outlined below, affordance-based reasoning may be the basis

for all manner of creative reasoning and sense-making.

4.7.1 Novel Tool Use

Consider the example of a robotic assistant helping a human with an assembly task in which the human has asked the robot to tighten a loose screw. We would like for the robot to understand this task and the tools needed from an intuitive standpoint such that even in the absence of a screwdriver, it can reason through alternatives and find another substitute.

The robot may know of a number of rules related to its role as a helper. One rule may be: that if agent X is given a task to tighten a flat head screw S, and X sees an object O that has a flat-head edge, then the object O has a $tightenWith$ affordance. This rule can then be represented in DS-theoretic uncertain logic a follows:

$r^0_{[\alpha_{R_0},\beta_{R_0}]} :\equiv$
$hasFlatEdge(O) \wedge task(X, tighten(S, flat)) \implies$
$tightenWith(S, O)$

The robot (through attention control and perceptual semantics analysis in CALyX) can look around the room and determine (within a certain uncertainty interval) whether or not each of the various objects that it sees has a flat edge.

$hasFlatEdge(Screwdriver)[0.95, 0.95]$
$hasFlatEdge(Knife)[0.9, 0.9]$
$hasFlatEdge(Coin)[0.75, 0.95]$
$hasFlatEdge(Pencil)[0, 0.95]$

The robot can then apply DS-theoretic logical inference on rules, such as the one above, and infer uncertainties for the $tightenWith(S, O)$ affordance for each of the five objects. Based on this inference, the robot can deduce that knives and coins can be used to tighten screws in the absence of screwdrivers, but pencils cannot.

4.7.2 Creative Problem Solving

An affordance-based approach might shed light on insight and creative problem solving scenarios that require an ability to think about a problem from a different angle, [MB14], or in our case, a different context. Affordance-based creative reasoning approaches are not new and have been attempted by Olteteanu [OF14]. However, these approaches are limited for the same reasons as others in the affordance literature, in that they cannot account for complex affordances in different contextual circumstances. We believe that an affordance representation of the form presented in this work may assist in modeling both creative and commonsense reasoning processes more effectively. Moreover, when coupled with mental simulation engines, the agent need not physically actualize the affordances in order to see their effects, choosing instead to simulate mentally in a suitable physics engine.

4.7.3 Role of Affordances in Sense-making

Our perception of affordances in our environment enable us to not only know what we can do with objects around us, but they also serve to tell a story about our current situation. For example, chairs and tables in a restaurant allow people to sit and eat their food. However, a collection of chairs without any tables in the middle of the restaurant would strike us as a bit unusual. Our need to make sense of the situation drives us to dig deeper and learn more about the reasons why there are no tables. This same need is what allows us to discover problems when there is a mismatch between what we see and what we expect to see. Reasoning about cognitive affordances in a more general way, as outlined in this chapter, has the potential to assist in such sense-making, which can be useful for artificial agents navigating in the open world.

4.8 Conclusion

As part of their interview process, many modern technology companies show prospective candidates an object they have never seen before and ask them to describe what

they think is the object's function. The purpose of the question is to test the candidate and probe their intellect to identify candidates with strong mental representations of affordance. Clever answers are often rewarded and stand as an example of human creativity. The ultimate goal of our research is to endow robots with the ability to find creative ways to use and manipulate objects and their environment.

In this work, we took the first steps towards our goal and proposed a novel computational framework based on Dempster-Shafer (DS) theory for inferring cognitive affordances. We demonstrated how our framework can handle uncertainties and be extended to include the continuous and dynamic nature of real-world situations. We believe that this, much richer level of affordance representation is needed to allow artificial agents to be adaptable to novel open-world scenarios.

Part II

Learning Knowledge – Case of Norms

Chapter 5

Learning Social Norms from Natural Language

In Part II of this dissertation, we turn our attention to one specific type of knowledge needed for sense-making – normative – that we saw (in Part I) was crucial to language understanding and affordance perception. In this part, we will explore how norms can be learned from natural language, how they are represented and how they can be inferred from behavioral observations. We will also explore how legal theory around the notion of "consent," which is a particular class of social norms, can inform research in human-AI interaction.

First, in this chapter, we pick up where we left of in Chapter **??** and consider how social norms associated with object affordances can be learned from natural language instruction.

5.1 Introduction

Using and manipulating objects in the environment requires a cognitive ability to perceive and evaluate their meaning, applicability and usefulness in relation to our own abilities to take action. Such a relational notion, known as an *affordance*, links action and behavior possibilities with objects and features present in the environment, enabling the ability to guide our behavior [Gib79, ZHL+17]. In robotics and

AI, affordances have served as the underlying theory for action perception and have been modeled using relational and machine learning techniques such as Bayesian Networks, Markov-Logic Networks, Conditional Random Fields, and Reinforcement Learning [Ste02, MLBSV07, UNSO15, KS16b, SM17]. While much of the affordance literature in robotics has focused on object or environmental affordances, some have considered "social affordances" and offered an approach to perceiving visual cues offered by social scenarios involving other agents (e.g., raised arm signaling a high-fiving affordance) [SRZ16]. However, these methods do not allow for socio-contextual dependency on *object* affordances.

[SS16b, SS18a] proposed the theory of cognitive affordances to address this problem. The theory of cognitive affordances uses a probabilistic-logic based approach capable of inferring affordances in the face of changing contexts, social norms, and epistemic uncertainty , i.e., it accounts for those object affordances influenced by factors beyond perceptual cues on the objects themselves ("cognitive affordances"). However, the question of how exactly to *learn* these cognitive affordances and *utilize* them on a robot is still open. The problem is especially difficult because these sorts of socio-contextual dependencies are, for humans, learned through a few exposures or instructions, and not through numerous trials and errors.

In this chapter, we address these open questions directly and propose two novel contributions: (1) a grounding and integration of cognitive affordance representation within a cognitive robotic architecture, and (2) an approach to learning these cognitive affordances from natural language instruction in the presence of epistemic uncertainty. The proposed approach allows for encoding, learning and immediately actualizing of a broad class of normatively-charged cognitive affordances, accounting for aspects of objects that the agent can directly perceive (e.g., object features) and aspects that are not self-evident or directly perceivable from the object itself (e.g., context and social convention associated with the object, goals of the agent).

We will use a kitchen-helper robot from [SS18a] as our running example, with the robot learning, from instruction, how to properly grasp a knife when using it and when handing it over to someone (at the blade). Although the proposed approach

is not limited to this particular example, or even embodied robotic systems for that matter, a concrete example of this sort will help tie it to past work and explain various aspects of the representation, inference algorithm, learning approaches and integration with a cognitive robotic architecture, to allow for normative behavior capabilities.

5.2 Theory of Cognitive Affordances

We are interested in the class of *affordances* that possess additional properties and dimensions beyond the simple Gibsonian notions (e.g., "sitability of a chair"). As noted earlier, this class of cognitive affordances is deeply influenced by contextual and normative factors including goals and intentions, prior knowledge and interpretations, ensemble scene information, mental state, experience and developmental state, social and moral conventions, and aesthetic considerations among others. We will build on a recent theoretical model of cognitive affordances proposed by [SS16b, SS18a] that represents affordances as condition-action rules (R) where the left-hand sides represent perceptual invariants (F) in the environment together with contextual information (C), and the right-hand sides represent affordances (A) actualizable by the agent in the situation (e.g., the rule that one should grab a knife by the handle when using it would be translated by specifying the grasping parameters as F, the task context of "using a knife" as C and the constrained grasping location together with other action parameters as A). Affordance rules (R) take the overall form $r \stackrel{\text{def}}{=} f \wedge c \underset{[\alpha,\beta]}{\Longrightarrow} a$, with $f \in F$, $c \in C$, $a \in A$, $r \in R$, and $[\alpha, \beta] \subseteq [0, 1]$. $[\alpha, \beta]$ is a confidence interval intended to capture the uncertainty associated with the truth of the affordance rule r such that if $\alpha = \beta = 1$ the rule is logically true, while $\alpha = 0$ and $\beta = 1$ assign maximum uncertainty to the rule. Similarly, each of the variables f and c also have confidence intervals associated with them, and are used for inferring affordances as described in more detail below. Thus, rules can then be applied for a given feature percept f in given context c to obtain the implied affordance a under uncertainty about f, c, and the extent to which they imply the presence of a.

Given a set of affordance rules, we can determine the subset of applicable rules by matching their left-hand sides given the current context and perceivable objects in the environment together with their confidence intervals, and then determine the confidences on the fused right-hand sides (in case there are multiple rules with the same right-hand side) based on the inference and fusion algorithm in [SS18a]. We will use the "confidence measure" λ defined by [NDS+13] to determine whether an inferred affordance should be realized and acted upon. For example, we could check the confidence of each affordance on its uncertainty interval $[\alpha_i, \beta_i]$: if $\lambda(\alpha_i, \beta_i) \leq \Lambda(c)$ (where $\Lambda(c)$ is an confidence threshold, possibly depending on context c), we do not have enough information to confidently accept the set of inferred affordances and can thus not confidently use the affordances to guide action. However, even in this case, it might be possible to pass on the most likely candidates to other parts of the integrated system. Conversely, if $\lambda(\alpha_i, \beta_i) > \Lambda(c)$, then we take the inferred affordance to be certain enough to use it for further processing.

From a systems standpoint, in order to process cognitive affordances, several functional units were proposed by [SS18a]. During inference, the functional units are meant to search through all available affordance rules of the form specified above in the agent's long term memory and populate a working memory with the relevant rules. Once the rules are in the working memory, the system can use these rules as the basis for perception and inference. An example cognitive affordance rule instantiation in this past work had the form,

$$r \stackrel{\text{def}}{=} hasSharpEdge(O) \wedge domain(X, kitchen) \underset{[0.8,1]}{\Longrightarrow} cutWith(X, O).$$

While this past work presented some crucial early theoretical foundations for using and performing inference with cognitive affordances, it was missing two key components. First, the past work did not suggest how these rules could could be grounded in a robotic system. For example, [SS18a] state that the results from affordance inference are "passed to the robot's action management system," but

they do not discuss how exactly this interaction might work and how an action management system might be able to use this information in connection with its own action repetoire and action knowledge. Thus, an open question is how can an agent use $cutWith(X, O)$, and what exactly do the predicates and variables in the logical representation mean in a robotic architecture. In this chapter, we describe such a grounding for an exemplary architecture and provide a grounded rule representation consistent with the cognitive affordance theory, but also tightly integrated with the robot's actuation and perceptual systems. In doing so, we will also need to revisit and modify the above-mentioned cognitive affordance rule example to tie the predicates in the rule representation to perceptual and action knowledge actually available in the system as well as contextual knowledge associated with the task the agent is performing.

Moreover, while [SS18a] have outlined an approach for performing inference with cognitive affordance rules, it is still an open problem as to how these rules might be learned. Here, we propose a solution based on learning from instruction, which at times, might be the only option available to an agent, for example, in situations where the agent does not have enough time to observe or if the agent is not able to collect enough observational data. Recent work by [SKO+17] discusses an approach for learning percepts and actions from instruction. Here, we propose extending this approach for learning not only perceptual and action predicates, but the rules themselves. By combining these ideas from past work, we provide a novel approach for learning normatively-guided affordances from natural language.

5.3 Grounding and Learning

To choose and manipulate everyday objects in socially-appropriate and context-dependent ways, we claim that *any cognitive system will require mechanisms for learning, representing and immediately applying arbitrary socio-contextual rules associated with these objects.* While the cognitive affordance theory provides a suitable rule representation, the rules must be grounded within the cognitive system (Section

5.3.1). Action management components must be able to guide perceptual and action components to check if a rule applies, and then apply the rule by constraining action choices and parameters, all under conditions of epistemic uncertainty (Section 5.3.3). Moreover, much of these social norms are conveyed via natural language. So the natural language components must be equipped to parse speech into the grounded rule representations (Section 5.3.2).

5.3.1 Enabling Affordance Processing in a Cognitive Robotic Architecture

To integrate affordance processing into a cognitive robotic architecture, we developed a separate component for maintaining the affordance rules and the inference algorithm for DIARC, an example architecture [SSKA07]. In addition we updated several components of the architecture to be able to handle the new types of information enabled by the new affordance component. This grounding within the architecture gives the affordance reasoning mechanisms described in Section 5.2 a concrete medium through which new rules can be added dynamically based on an agent's interactions with its environment. This extends the functionality and utility of the theoretical model which previously was limited to a fixed set of abstract rules.

We selected DIARC over other cognitive robotic architectures (e.g., SOAR [LNR87], ACT-R [ABB+04]) because of its integration of social behaviors, and specifically its natural language understanding and production capabilities which allow for more natural human-robot interaction, as well as the ability to learn new concepts through natural language [SKO+17]. None of the current cognitive robotic architectures (including DIARC) are currently able to represent and reason about cognitive affordances.

So regardless of the choice of architecture, an affordance component of the type described here could be desirable to enable affordance processing and enhanced social interaction capabilities. Whichever architecture is chosen, the proposed affordance component will still need to be connected to other high and low level components in order to be able to influence perception and action.

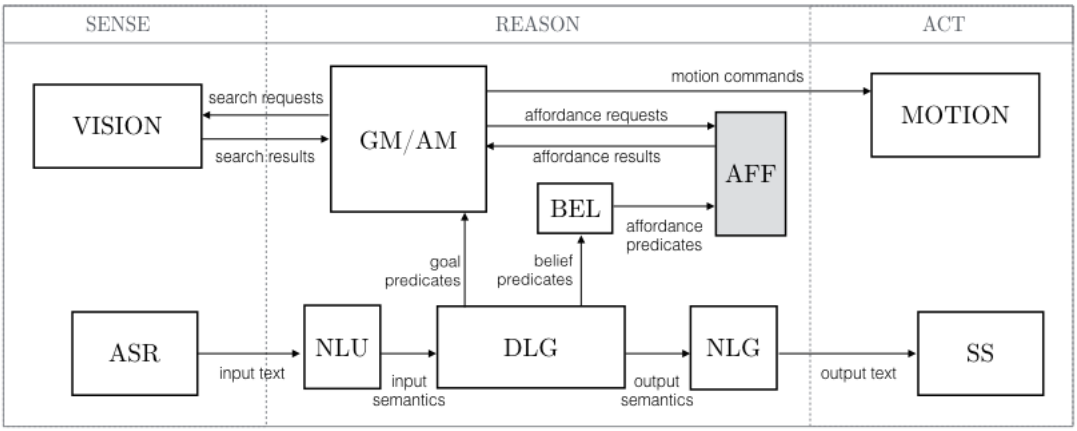

Figure 5.1: Architecture diagram. Affordance Component AFF highlighted. Other relevant components: Speech Recognition ASR, Natural Language Understanding NLU, Dialogue Manager DLG, Natural Language Generation NLG, and Speech Synthesis SS, Belief Model BEL, Motion Control MOTION, Vision VISION, and Goal Manager/Action Manager GM/AM. During operation, AFF receives semantic information, uses GM/AM to direct VISION to look for environmental features relevant to social norms, and then guides MOTION via GM to perform a socially-appropriate action.

Fig. 5.1 depicts the integration of the affordance component (AFF or AFFORDANCE) in DIARC. The subcomponents of AFFORDANCE work closely with sensory and perceptual systems (e.g., vision system) and other components in the architecture to coordinate perceptual and action processing. AFFORDANCE is connected to the goal management component (GM/AM or GOAL MANAGER), and during the execution of actions GOAL MANAGER sends affordance requests to AFFORDANCE. These requests provide information about the current action to be performed and the context. AFFORDANCE returns the specific perceptual features that need to be searched in the environment. This allows GOAL MANAGER to direct the attention of low-level perceptual systems like the vision system (VISION) to perform searches in a focused manner, only looking for perceptual features relevant to the applicable rules in AFFORDANCE. The presence or absence of the searched perceptual features (along with perceptual uncertainty information) is passed back to AFFORDANCE, which subsequently performs uncertain logical inferences (logical AND and modus ponens) on the rules.

In dialogue-driven tasks, GOAL MANAGER receives language based goals via

the natural language pipeline (ASR → NLU → DIALOGUE), while the belief component (BEL or BELIEF) is the recipient of language-based knowledge. BELIEF maintains a history of all declarative knowledge passing through the architecture and is capable of performing various logically-driven knowledge-representation and inference tasks. Thus, it serves as a convenient holding-area for cognitive affordance information partially processed through the natural language pipeline, which can then be retrieved and processed by AFFORDANCE.

Integrating AFFORDANCE into DIARC, or any cognitive architecture for that matter, requires more than simply depositing it into the system. Various other components (GOAL MANAGER, VISION, BELIEF, etc.) must also be modified so they can provide the additional capabilities required for cognitive affordance processing. For example, the natural language pipeline must be updated to allow for the recognition and understanding of cognitive affordance related words ("cutting", "sitting","enclosing", etc.), and GOAL MANAGER must be updated to recognize these semantic representations and consult AFFORDANCE at the appropriate points in action selection and execution. In the next sections, we will discuss in more detail the specific architectural modifications, and their resulting functionality which enables these new operations.

As an additional benefit, these modifications provide the architecture with the new ability to understand references to objects by their affordance (e.g., a knife as not just an object with some visual property, but as an "object used for cutting"). An affordance-enabled cognitive robotic architecture also allows an agent to account for context, and helps constrain and guide behavior. Actions no longer need to be performed the same way each time, but can vary depending on context (grasping a knife can be done differently depending on what the context requires). For example a kitchen helper robot may grab a knife differently if the context of the grab is that the robot will use it to cut something, as opposed to the robot grabbing it so it can be handed to a human. Or it might carry plates of food differently in the context of serving them versus the context of busing a table.

5.3.2 Learning Affordance Rules from Instruction

As mentioned earlier, we will use as our simple guiding example of the instructions "a knife is often used for cutting" and "to pickup a knife grab it by the handle" given to a robot that does not know an epistemically uncertain ("often") functional affordance ("cutting") of a knife, or its context-sensitive ("pick up") grasp affordance ("by the handle").

We previously outlined how the affordance model described in Section 5.2 can be integrated into a cognitive robotic architecture to expand the capabilities of the robot. This integration also gives that model a mechanism through which new rules can be added on the fly, allowing it to better adapt to and represent real world scenarios. In order to do this, various DIARC components must be extended. Natural language utterances that contain cognitive affordance rules must be converted to general purpose facts stored in BELIEF, and then used by AFFORDANCE to generate the rules described in Section 5.2, which can then ultimately be used to perform uncertain logical inference. With these extensions to existing DIARC components we are able to leverage DIARC's mechanisms for learning through instruction (see [SKO+17]) to enable learning new affordance rules about concepts that the agent already *understands* as well as completely novel concepts which have been learned on the fly.

The role of the Natural Language Understanding component is converting the text form of spoken utterances into a semantic form which can be *understood* (used) by the other components within DIARC. We extended this component through the addition of new parsing and pragmatic inference rules which enable the generation of new semantic forms.

In order to learn from natural language instructions, a cognitive robotic architecture must be able to ground the content of the utterances containing the instructions in terms that it is able to understand. AFFORDANCE *understands* affordance

rule descriptions which are represented in the predicate form,

$$implies(antecedents, consequents, confidence),$$

where the predicate's arguments represent the antecedents, consequents, and confidence interval of an affordance rule. The natural language processing components of DIARC (See Figure 5.1 for ASR, NLU, DM) convert spoken language into a predicate of this form and assert it into BELIEF.

When an utterance is spoken to the agent, the speech recognition component (ASR) converts the acoustic speech signals to text. The natural language understanding component (NLU) receives the utterance in text form from the speech recognition component and performs two steps of processing. The first step parses the text into a form that can be used by the rest of the system. The second performs pragmatic inference to add a notion of the speaker's intent to the representation of the utterance [SBC+13].

In the parsing step, the natural language understanding component uses a parser to determine the syntactic structure and the semantic interpretation of the utterance. The parser used in this configuration of DIARC is an extended incremental version of the Combinatory Categorial Grammar (CCG) parser from [DSBS09], described in more depth by [SKO+17]. It contains a dictionary of parsing rules each composed of three parts: a lexical entry, a syntactic definition in CCG, and a semantic definition in lambda calculus. An example set of rules can be found in Table 5.1. These rules are a subset of the complete set of rules used by the system. They are selected because of their relevance to the empirical demonstration in Section 5.4.

An example of a cognitive affordance rule spoken in natural language and its accompanying semantics are,

"To pickup a knife grab it by the handle",

Table 5.1: A subset of the relevant rules used by the natural language understanding component (NLU).

Label	Syntax	Semantics
to	(S/C)/C	$\lambda x \lambda y.implies(x, y, high)$
pickup	C/NP	$\lambda x.pickUp(?ACTOR, x)$
a	NP/N	$\lambda x.x$
knife	N	$knife$
grasp	C/NP	$\lambda x.grasp(?ACTOR, x)$
the	NP/N	$\lambda x.x$
by	(NP/NP)\NP	$\lambda x \lambda y.partOf(x, y)$
handle	N	$handle$

$$STATEMENT(Sam, self, implies(pickUp(self, knife), graspObject(self, partOf(handle, knife)), high)).$$

Here, "Sam" is the name of the human (and trusted source) speaking to the robot. This representation denotes a statement from Sam to the agent whose semantics are the *implies* predicate, above.

The parsing step produces a notion of the meaning of the spoken utterance. The pragmatic inference step uses that meaning and a set of inference rules to determine the speaker's intention. The pragmatic inference system used in our configuration of DIARC is described in work by [SBC+13]. In the case of our working example the semantic representation generated in the parsing step matches the left hand side of the rule,

$$STATEMENT(A, B, X) \implies wantBelieve(A, B, X),$$

which is a general rule for utterances of the type *STATEMENT*, and can be interpreted as "when a person tells an agent something it wants the agent to believe it". The resulting DIARC representation produced by the natural language understanding component (NLU) is the predicate,

$$wantBelieve(Sam, self, implies(pickUp(self, knife), graspObject(self, partOf(handle, knife)), high)).$$

This semantic representation from NLU is received by DIALOGUE whose role is to appropriately respond to utterances from other agents. In the case of our example DLG recognizes that Sam wants the agent believe a predicate. It checks if Sam is a trusted source of information, and if so, asserts the predicate into BELIEF. Upon confirmation that the information has been successfully stored, DIALOGUE submits a goal to GOAL MANAGER to verbally acknowledge understanding.

Once the information from the utterance has been asserted into BELIEF it is accessible to AFFORDANCE. Epistemically, an utterance is a piece of evidence received by the agent in support of the truth of the affordance rule it represents. Thus, we use the confidence directly to represent the degree of support for the rule.[1] The confidence value may be used to capture the inherent uncertainty in the utterance (e.g., when qualifiers such as "sometimes" or "maybe" are used), or the trust placed in the interlocutor (e.g., a rule taught by a superior or boss may hold more water), or the uncertainty in speech detection mechanisms, or in some combination of these factors. The linguistic placeholder "high" represents a preset confidence (0.95). That value is used because the speaker is a priori known to be trustworthy by the agent, but nothing in our system requires this particular, method of assigning confidence values.

Functional affordances can be learned in the same way as the action affordance describe above. For example, the utterance "A knife is sometimes used for cutting" would be translated to the DIARC predicate representation $implies(knife, cutting, mediumLow)$ in BELIEF.

[SKO+17] describes how DIARC agents can learn new concepts on the fly through natural language instruction. When the agent encounters an unknown word it is able to infer its syntax and semantics based on the parser's knowledge about the syntax and semantics. In the case of utterances related to cognitive affordance

[1] The "confidence" here is different from the confidence measure λ discussed in Section 5.2. λ is a singular measure of the degree of uncertainty of an uncertainty interval (somewhat akin to the width of the interval) typically used in conjunction with Dempster-Shafer theory of uncertainty. We can use λ when executing affordance-based commands (Section 5.3.3) and deciding which action to perform when there are multiple choices. However, the confidence value mentioned here is used to directly represent the degree of support for rule, i.e., it represents the single-valued precision assigned to the rule when received as evidence via an utterance.

inference rules, the syntax and semantics of previously unknown antecedents or consequents can be inferred by recognizing the pattern of the rest of the utterance. Novel consequents or antecedents introduced this way can be recognized in subsequent utterances and their representation in the set of rules in AFFORDANCE will be consistent. This enables the agent to understand cognitive affordance rules with previously unknown consequents and antecedents, which provides the agent the ability to continuously adapt its knowledge base.

To clarify it is worth noting that the architecture proposed by [SKO+17] was limited to learning concepts that have direct perceptual correlates (speech signals or visual attributes), and was not able to learn and utilize non-perceptual or cognitive concepts (like cognitive affordances). These involve non-perceivable attributes (contexts), and relationships between agent capabilities (actions) and perceptual entities (visual features) all tied together in compact natural language utterances. In the next section, we describe how an agent having learned cognitive affordance rules can apply this knowledge immediately in a command-based task.

5.3.3 Executing Affordance-Based Commands

There are numerous examples of DIARC and other cognitive robotic architectures enabling robots to engage in task based dialogues where a human is able to instruct a robot to perform tasks using commands given in natural language. The integration of AFFORDANCE into such architectures allows for the incorporation of affordance information when discussing a task. This allows for a more natural dialogue, and gives the human and robot more flexibility in the objects they discuss.

Returning to our running example of using a knife, consider a human uttering a command to the robot: "Pass me something used for cutting." Currently, DIARC would fail because the GOAL MANAGER would not be able to handle pairing a known action of "passing" with a non-specific object reference "something" and a functional affordance concept of "cutting." Moreover, even a more specific request of "pass me a knife" would often fail, because there is no guarantee that the robot will choose to pass the knife by grasping the blade (the normatively appropriate option), as opposed

Figure 5.2: Left: knife; Right: Grasp candidates all across the knife. Cognitive affordances can serve as a normative constraint when selecting one of these many possible grasp possibilities.

to the handle, which has similar – if not better – grasp possibilities. As shown in Figure 5.2 (taken from [SKO+17]), there are many available grasp candidates distributed all across the knife on the handle and on the blade.

In order to *understand* the command, the information contained in it needs to be grounded within the system. We start with the system knowing nothing about knives or how to pass them. We use the features of DIARC described in [SKO+17] to teach the system what a knife is and how to pass something. At this point, the robot knows that an observed 3D point cloud is a "knife" and that certain subsets of this point cloud constitute "handle" and "blade". Using the object grasping mechanism described in [TP14] it is capable of generating candidate grasp points (from the geometry of the point cloud) and then scoring these grasp points to determine which ones are likely to succeed. We use a four-layer deep convolutional neural network to make grasp predictions based on projections of antipodal grasp points contained between fingers.

Using the approach described in Section 5.3.2, we assume that a human has taught the robot cognitive affordance rules about a knife in three utterances as follows,

"A knife is used for cutting",

"To pickup a knife grab the knife by the handle",

"To pass a knife grab the knife by the blade."

The following predicates are produced in BELIEF, as described earlier,

$implies(knife, cutting, high)$,

$implies(pickUp(self, knife), graspObject(self, partOf(handle, knife)), high)$,

$implies(pass(self, knife), graspObject(self, partOf(blade, knife)), high)$.

Now that the system *understands* how to pass knives in the context of cutting we can instruct it to do so using the natural language mechanisms described earlier,

1. Utterance: “pass me something used for cutting"
2. Parse: $INSTRUCT(Sam, self, pass(self, usedFor(something, cutting)))$
3. Relevant Pragmatic Rule:

   $INSTRUCT(A, B, X) \implies want(A, X)$
4. DIARC Semantic Representation:

   $want(Sam, pass(self, usedFor(something, cutting)))$
5. Submitted Goal Predicate:

   $pass(self, usedFor(something, cutting))$

Upon goal submission GOAL MANAGER executes the action script associated with the goal. An action script is hierarchically organized with actions and sub-actions, with bottom-level actions representing commands issued to the action component (MOTION). The hierarchy for the “pass” action is shown in Figure 5.3. Executing an action script of this form involves performing a preorder traversal of its tree. At each node, we perform three operations for applying learned affordances, in addition to the operation related to the action itself.

First, an affordance request is sent to AFFORDANCE to $getFeatures()$, which involves assimilating newly learned affordance rules, identifying relevant affordance rules and returning perceptual invariants (F) from the antecedents.

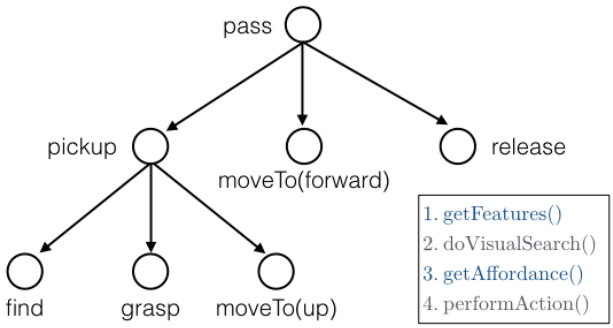

Figure 5.3: Key affordance-related operations during action execution using an exemplary *pass* action script. For every node in the action tree four operations are performed: extracting perceptual features and contextual items from the relevant affordance rules, running a visual search to determine whether these features exist in the agent's environment, performing inference with the rules and observations to obtain constraints on actions, and performing the action with inferred constraints.

AFFORDANCE queries BELIEF for any new $implies(X, Y, Z)$ predicates it may have learned since its last call. AFFORDANCE maps the arguments into the perceptual invariants (F), contextual items (C) and affordances (A) in the affordance model. Here, we assume that context itself is a higher-order action and therefore is captured as a functor-name in the DIARC semantic representation of the utterance. Thus, "pass" is the action context in $implies(pass(self, knife), graspObject(self, partOf(handle, knife))\ high)$. We recognize that context is not always knowable or definable in advance, but in this situation, contextual information is explicitly provided in the utterance, and is therefore available for the system to use. In other instances, the context may be more implicit and the agent may need to infer it; the proposed approach does not preclude this sort of inference because the affordance model is generalizable to capture context predicts regardless of how they are obtained, and that is the subject of future work. The perceptual invariant (F) is available in the argument to the action context and thus, for example, "knife" is a perceptual invariant to be added to the rule. The affordance information (A) is available from consequents where the "graspObject" predicate is flattened. This process, leads to the following affordance rules in our motivating example,

$$r^1 \stackrel{\text{def}}{=} knife(K) \underset{[0.95,1]}{\Longrightarrow} findObject(cutting, knife(K)),$$

$r^2 \stackrel{\text{def}}{=} knife(K) \wedge context(C = pickup) \underset{[0.95,1]}{\Longrightarrow}$
$graspObject(knife(K), handle(P), partOf(P, K)),$

$r^3 \stackrel{\text{def}}{=} knife(K) \wedge context(C = pass) \underset{[0.95,1]}{\Longrightarrow}$
$graspObject(knife(K), blade(P), partOf(P, K)).$

Once the rules have been updated to include new additions from BELIEF, AFFORDANCE selects the rules relevant to the current situation. While we do not provide an in-depth comparison of various rule-selection approaches, we take a straightforward approach in selecting those rules relevant to the current action (i.e., the current active node in the action tree) and affordance. We select rules with consequents containing functor names that match the current action. This is possible since the syntax and semantics of the affordance predicates match the grounded representations of actions in GOAL MANAGER and MOTION. In addition to the current action, we use goal predicate information (including affordance) obtained from the current command to further prune the rules, if necessary. AFFORDANCE obtains this information by querying BELIEF for *usedFor*(X, Y).

During the "find" action, the only match is rule r^1 and thus the only rule that is selected is associated with functional affordance of the knife. The output from *getFeatures*() is then sent as a search request by GOAL MANAGER to VISION to locate them in the agent's visual field of view. For example, the perceptual invariant ($knife(K)$) obtained in the "find" action is then sent back to GOAL MANAGER, which then initiates a visual search to look for a knife. Upon finding a match, VISION provides a detection confidence for the object it has identified as being a knife and AFFORDANCE uses this confidence to perform inference to determine if the deduced action "find" is above a certain confidence threshold. If so, the object identified as a knife will be selected for further processing.

If a representative visual object is identified, then a second request is made to the affordance component to *getAffordance*(), during which affordance inference is performed and the best action or object constraint is returned. The constraints are then used in connection with the motion commands and sent to MOTION. Thus,

instead of the general command to grasp a knife, which could result in the agent selecting one amongst a countless number of high-scoring grasp candidates on the knife, the agent may be constrained to only selecting those on the handle.

At the next node (grasp), this process is repeated, but now there are two rules associated with grasp, however only one is associated with the context of "pass." Thus, inference is performed on this one rule and the constraints $knife(K)$, $blade(P)$, $partOf(P, K)$ is returned and used for identifying grasp points on the blade of the knife. The agent can then correctly (from a normative standpoint) passes the knife by the blade.

5.4 Evaluation

We take a two-step approach to evaluating the affordance-enabled cognitive robotic architecture. First, we evaluate whether the system implements a correct algorithm (i.e., is it taking the correct actions when instructed with a cognitive affordance-based natural language command?). We do so through an extended simulation, in which the entirety of the architecture was put to the test without external noise or sensor fluctuations that typically occur in real-world settings. Clearly, empirical real-world runs of the system are important and show how the architecture can perform, not just in simulation, but on an embodied robot, in which real-time constraints apply. So, in our second step of the evaluation we tested the architecture on a PR2 robot and provide an uncut video. We describe each step of the evaluation below.

5.4.1 Simulation Experiment and Empirical Demonstration

We first tested the correctness of the approach in an extended simulation involving several household objects and over two-dozen rules. As noted earlier, the goal was to be able to test the proposed approach in a simulation without perceptual and sensory noise experienced in the real world in order to focus on evaluating the correctness of the various underlying algorithms.

For the experiment we considered the following eight household objects, each

composed of two parts: Knife (handle, blade), spatula (handle, blade), spoon (handle, bowl), shoe (upper, sole), hammer (handle, head), glass (bowl, stem), mug (handle, barrel), screwdriver (handle, shaft). We considered five different affordances in the spirit of those used in computer-vision datasets associated with affordance detection [MTFA15, Var15]: containing, cutting, pounding, rolling, and poking.

We restricted the agent's action repertoire to the actions *pass*, *pickUp*, and *pointTo*. Unlike *pass* and *pickUp*, *pointTo* involves finding but no grasping. With these objects and actions, we generated 15 different commands (5 affordances times 3 actions) of the form "[action] something used for [affordance]" (e.g., point to something used for pounding).

We are interested in learning functional affordances and action affordances that contain a notation of confidence. To capture this, we used four terms to represent different degrees of confidence: occasionally, sometimes, often, generally, which we then mapped to specific numerical values [KC15]. Thus, given eight objects, five affordances and four uncertainties and three actions, we can generate 288 possible affordance rules (160 functional affordance rules and 128 action affordance rules).

In any given learning scenario, the agent is taught a set of rules chosen from these 288 possible rules, thus generating 2^{288} different possible learning scenarios. Also, since we have eight objects, we can generate 256 possible scenes involving these objects, which combined with the 15 possible commands we can give the agent 3840 different problem situations (scene-command combinations). Testing the proposed architecture across all these possible learning scenarios ($3840 * 2^{288}$) and problem situations is infeasible.

Instead, we evaluate the system by (1) choosing a random subset of our evaluation space, and (2) establishing some general performance expectation for the system in this evaluation space. We limit our evaluation space by randomly choosing 10 different scenes and testing the agent's performance for all 15 commands. With regards to the selection of rules, we arbitrarily choose two different rule sets representing two different normative standards and provide some expectations for how we expect the agent to act based on these two distinct learning scenarios. In

Scenario A, for each of the five affordances, we generated a list of objects (ranked highest to lowest confidence) that possess this affordance. In Scenario B, we reversed the ranking of objects. For example, in Scenario A, a mug is top-ranked object for the containing affordance while a shoe is a bottom ranked (but still feasible) object. While, in Scenario B, the shoe is top-ranked and the mug is bottom ranked. These two scenarios represent our ground-truth rules and Table 5.2 shows these values.

Table 5.2: Ground Truth rules in Scenario A. Scenario B can be obtained by reversing the list of each affordance. Point confidences in parentheses.

Affordance	**Generally (0.95)**	**Often (0.75)**	**Sometimes (0.5)**	**Occasionally (0.25)**
containing	[mug, glass]		spoon	shoe
cutting	knife		screwdriver	[spatula,spoon]
pounding	hammer	shoe	spatula	mug
rolling		glass	screwdriver	
poking		screwdriver	knife	

We used Table 5.2 to derive functional affordance rules of the form "[object] is [uncertainty] used for [affordance]" (e.g., "a spatula is sometimes used for pounding"). For action affordances, we generated 16 rules corresponding to physical grasp affordance rules in the pass and pickup context for all eight objects, with a single confidence setting of "generally." For example, "To pass a shoe, generally grab the shoe by the sole." Of the 288 rules initially stated, many are somewhat nonsensical by our own normative or practical standards (e.g., a mug being used for cutting). However, the robot does not know this, and we therefore expect the robot to be able to perform the necessary affordance inference without this additional commonsense knowledge.

We performed multiple trials during which we generated various tabletop scenes using combinations of the eight objects. We tested both learning sets of affordance rules, and acting on sets of commands using the proposed architecture. We evaluated the performance of the system by checking if the following principles held true in all trials: (1) if all the objects are in the scene then the robot must select the top-ranked object for the required affordance, (2) if the top-ranked object is not available, then the robot must select the next lower ranked object, (3) if there is more

than one top-ranked object with equal measures of confidence, then the robot may select either, and (4) if there are no objects available with the required affordance, the robot must tolerate the failure condition and provide a suitable response.

We ran the experiment across 10 randomly generated scenes of varying sizes (including one with all the objects). During each run, we taught the 32 above-mentioned rules in each of Scenario A and B, then we presented a randomly generated scene, and issued each of the 15 commands in sequence. We ensured that set of scenes in combination with the commands covered the above-mentioned four performance expectations. Table 5.3 shows the general form of the two types of affordance related utterances (Utterance Templates) our system can handle, and the component parts of those templates that can be expanded as needed for what ever application the system is being used for (Grounded Concepts), provided they can be grounded in the architecture. It is important to note that these utterance types are not the only language DIARC can understand, they are *added functionality* that coexist with prior functionality.

Table 5.3: High level syntax of understandable utterances, in JSpeech Grammar Format (JSGF).

Utterance Templates	
<statement >	a <object> is [<qualifier>] <implies> <use> | to <goal> a <object> [<qualifier>] <primitive> <object> <mod> <part>
<command>	[now | okay | first | then] (<goal> | <primitive>) something <implies> <use> | <primitive> <object> <mod> <part>
Grounded Concepts	
<qualifier >	sometimes | often | generally | always
<object >	mug | knife | wine glass | spatula | spoon | shoe | screwdriver | rock
<implies>	used for
<primitive>	grab | grasp
<goal>	pass | pickup| point to
<mod>	by the
<part>	red | green | blue
<use>	cutting | containing | pounding | rolling

In this evaluation we are interested in whether the integrated system correctly learned the rules from natural language expressions, and then immediately applied this knowledge correctly to select the best action. The two learning scenarios described above provide a ground-truth of sorts as the objects are ranked from best

to worst in each scenario. It is important to note here that we are not interested in evaluating robustness of the underlying low-level perceptual and action systems themselves. Accordingly, we avoid sensor noise and motion imperfections by running this experiment as a simulation, and focus exclusively on evaluating the proposed architecture with AFFORDANCE. Moreover, the rule uncertainties were set to four distinct and separated values to ensure that the ground-truth rule sets themselves were not noisy, i.e., without overlapping uncertainty intervals. Since we are evaluating a normative system, we lose the ability to clearly establish a ground truth if the underlying rules themselves were noisy. For example, if the uncertainties of a knife and screwdriver as objects used for cutting are very close to one another and overlapping, then which object is a better object becomes a more difficult question and without a clear answer supported in the ground truth. Thus, given a clear ground-truth and no sensory noise, the proposed architecture should learn the rules and act correctly 100% of the time.

As expected, we obtained a 100% success rate with the robot inferring the correct functional affordance and choosing the correct object (for all actions) and choosing the correct grasp locations (for pass and pick). We observed this success rate across all scenes measured. As one example, when all the objects were presented, the robot chose the mug when asked to select an object with the containing affordance. Likewise, the robot correctly identified top-ranked objects for each of the four affordances. This meant that for Scenario B, the robot correctly identified the shoe as being the best candidate with a containing affordance.

Our simulation further suggests that any non-100% performance must be due to sensor noise. If the agent is unable to correctly detect that an object on the table is in fact a knife, when asked for something used for cutting, then the agent is less likely to find this object as a suitable candidate - affordance inference will yield an uncertainty that might be below a threshold confidence measure described earlier.

In addition to the simulation, we further provide an empirical demonstration of a DIARC agent with a fully integrated cognitive affordance reasoning system in a task driven dialogue involving multiple human interlocutors. In this demonstration

we show the system's ability to learn new cognitive affordance rules on the fly, and to reason about these newly learned rules. A video of this demonstration is located here: **http://tiny.cc/affordanceNL2018**. We use the motivating example utterance "Pass me something used for cutting" spoken from a human to robot running cognitive affordance-enabled DIARC.

5.4.2 Commands with Implicit Affordances

Thus far, we have presented examples where the requested affordance was explicitly stated, such as "pass me something used for pounding." However, the approach presented in this chapter is not limited to such cases and is capable of handling cases where the requested affordance is not explicit. For example, a command "pound the nail" contains an implicit request for a tool that can do the pounding. In some sense, the command might actually be suggesting "pound the nail *with something used for pounding*," without saying so explicitly. As before, the robot is taught the normative affordance rule that a hammer can be used for pounding. But, in order for the robot to be able to make use of this affordance rule, it must already have an action script that describes how it should perform the pounding action. Much like the action script depicted as a tree in Figure 5.3, we consider an example action script for *pound* that is composed of a *pickup* action. It also contains a $moveAbove(nail)$ action and a series of repeated *raise* and *lower* actions to generate the pounding motion. The *pickup* sub-action, in turn, contains *find*, *grasp* and $moveTo(up)$ sub-sub-actions. The *pound* action can be made more complex containing visual actions of sensing the depth of the nail and identifying when to stop pounding.

In addition to this action description, the action script needs additional knowledge to handle cases when the action is called with a tool explicitly mentioned (e.g., pound the nail with the hammer) and when the action is called with the tool affordance implicitly suggested (e.g., pound the nail). In the implicit case, the highest level *pound* action must be able to supply a $usedFor(something, pounding)$ argument to the child *pickup* action. Thus, the *pound* action must first review the arguments of the goal predicate $pound(self, nail)$ received and parsed from the command

utterance and then provide the arguments *self* and $usedFor(something, pounding)$ to the the *pickup* action. Now, it is possible that the implicit command "pound the nail" was intended to be interpreted as "pound the nail *with the hammer*". In this case, the action script would need to consider other factors (e.g., the intent of the speaker) in order to determine what exactly was left unsaid – i.e., was it that the speaker intended for the agent to use a specific tool, namely the hammer, or any tool with a pounding affordance.

With this translation, the rest of command execution proceeds as described in the previous sections. This example shows that with suitable modifications to the action script, we can handle commands that contain affordance information explicitly as well as implicitly. It is important to reiterate a key assumption: that the robot is already equipped with the above-mentioned *pound* action script. Learning these action scripts (from natural language or however else) as well as determining interpretations of unsaid action arguments is beyond the scope of this chapter and subject of ongoing work.

5.5 Discussion

The above walk-through and simulation show how a set of new social norms, previously unknown to the agent can be acquired, in one-shot, from natural language instruction. The process of learning an implication rule of the form described is generalizable to other rules as long as the agent is familiar with the entities being described. That is, the agent already knows what a knife, handle, and blade mean. Critically, the new knowledge of the social norm encoded as an affordance rule is now available for inference by any and all subsystems in a cognitive robotic architecture. As shown in the evaluation, these rules are put to immediate use by means of a follow-up request from the human. These are, to our knowledge, the first demonstrations of an agent learning an unknown affordance norm from natural language instruction and then immediately performing an action sequence conforming with the rule that it just learned. Moreover, an affordance norm of this sort may be ben-

eficial to not just an action subsystem of an architecture, but to planning and other subsystems. A general rule-based structure, coupled with an inference mechanism presented here allows these other subsystem to query and access these affordance norms, as well.

Note that the above demonstration also shows that the instructions and actions do not have to pertain to a particular set of sensors or actuators and do not depend on a particular robotic platform. Rather the same inference and learning mechanisms can be carried out in other agents with different action capabilities. It is also important to note that the proposed approach and learning methodology are not limited to the particular examples demonstrated, but being implementations of a general framework for reasoning about affordances to guide normative behavior, are only limited by the agent's knowledge of natural language as well as its sensory and actuation capabilities (e.g., a Nao robot may not possess adequate gripper capabilities to grab a knife, but will still be able to reason about the normative aspects of other action capabilities like pointing and can still learn from instructions about these normative aspects of these actions).

Finally, to demonstrate the extent of learning, we note that current state of the art vision systems can identify and label objects and object features with a high level of accuracy. Thus, an agent can potentially become familiar with names and descriptions for thousands of objects. Along the same lines, agents can be trained through existing methodologies to build a substantial vocabulary and grammar allowing for descriptions of an infinite possible set of descriptors for perceptual invariants, contexts and actions. Hence, it is not possible, nor does it make sense, to evaluate the system exhaustively by generating every possible combination of rules and checking whether the agent can learn these rules. The strength of our system is that no mater what set of rules we give it, provided we have the sensory information to ground them, it will be able to learn and reason about affordances.

5.6 Related Work

While affordances have been studied for decades in philosophy and psychology, few computational approaches have been presented for modeling them in normative contexts, and none for learning them *from* natural language, which is an important open problem in affordance-related research in robotics [ZHL+17]. We believe that the proposed approach represents a significant advance over existing approaches. Existing work in cognitive robotics as well as in AI originated from the general philosophical and psychological theories and diverged in two directions: statistical approaches and ontological approaches. The statistical approaches modified and implemented these general theories in specific domains using statistical formalisms to represent and compute affordances [Ste02, MLBSV07, AMC14, CPCI15, UNSO15, KS16b]. The affordances were modeled as a statistical relationship between an object, actions performed on the object and the effects of those actions (i.e., success or failure). There has been some preliminary work to extend this approach by incorporating "environment" as a fourth entity, thereby providing some degree of situatedness and context [KSTN11]. The ontological approaches focused on developing a detailed knowledge-ontology based on conceptual, functional and part properties of objects, and then used a combination of detection and query matching algorithms to pinpoint the affordances for objects [Var15]. However, neither approach considered the influence of social or normative (and non-perceptual) factors in affordance perception.

More recently [SRZ16] presented a framework for reasoning about "social affordances" and provide a system that can act in social scenarios. However, the underlying affordance model is still largely devoid of contextual or normative reasoning, i.e., non-perceivable aspects of affordances, and is focused just on physical geometries of objects, i.e., perceivable aspects of affordances, in these scenarios (in this case skeletal geometries). Other work in robotics has explored mechanisms for detecting context and social contextual perception at both an individual level [OR15, NR15, PFS+15], as well as in group-level activities [OA14]. However, these approaches do not provide a generalized model or integration of normative affordance

perception of objects in a robotic architecture.

Thus, more generally, despite these past efforts, the task of computationally modeling affordances faces many challenges that have not been overcome in the previous work. These past efforts do not allow for reasoning about normative affordances, and from an architectural standpoint, most affordance processing is subsumed by the sensor processing (e.g., vision) or higher-level cognition (e.g., planning), which does not allow for an effective interaction between top-down and bottom-up processing of information in these past systems. Moreover, none of the current approaches show how affordances can be learned from natural language.

5.7 Conclusions and Future Work

The expressive framework of cognitive affordances treats an affordance as normative condition-action rule. In a sense, it extends the traditional Gibsonian notion of an affordance as a relation between an object and an action to include other non-perceptual aspects influencing action selection such as context, intentions, and social conventions. In this chapter, we provide two contributions: (1) a grounding and integration of this theoretical framework within a robotic architecture, and (2) an approach to learning cognitive affordances from natural language instruction. To accomplish this task, we extended recent work in instruction-based one-shot learning to be able to parse and learn cognitive affordance rules. The predicates and terms that constitute the rules contain perceptual and action concepts that are grounded within the DIARC cognitive robotic architecture. For each action that the robot must perform, we proposed several operations that obtain sensory information from the robot's perceptual system, perform inference over a set of relevant cognitive affordance rules that constrain the action, and execute the constrained action. We evaluated the approach through an extended simulation and empirical real-world runs of the proposed robotic architecture implemented on a PR2 robot. Critically, we were able to show that not only can an agent learn normative behavior from instruction, but immediately apply this newly acquired knowledge to the task at

hand. This to our knowledge is the first conceptual and robotic demonstration of an agent learning an unknown affordance norm from natural language instruction and then immediately performing an action sequence conforming to the rule that it just learned. We believe that these capabilities are necessary to allow agents to work effectively with humans and dynamically learn and perform tasks in a way that respects prevailing social norms. The approach presented in this chapter does not currently incorporate commonsense knowledge about objects and their similarity to other like-objects. Thus, the cognitive affordance rules that are learned from natural language are limited to the particular object explicitly taught. One direction of future work is to explore how to induce new cognitive affordance rules using commonsense knowledge. For example, we would like for the system to know and use the fact that knives and screwdrivers are both tools that have pointed ends and must be handled carefully. Then, when we teach the robot how to safely pass a knife, we would like for it to subsequently induce a comparable rule for the screwdriver, as well.

Chapter 6

Cognitive Science of Norms Representations

In this chapter, we take on the challenge of deepening our understanding of how humans mentally represent and learn norms and what characteristics of norms are captured in this representations. We then use these findings to inform a computational model for learning norms from behavioral observations.

6.1 Introduction and Motivation

Someone's cell phone begins to ring in the library. The person quickly answers it by whispering "hold on," then leaves the library and takes the call in a normal voice outside. The person understands that taking a phone call in the library is not socially acceptable, though briefly whispering is. Somehow, the situation activated a set of norms in this person's mind, including: "when someone calls you, you should answer the phone"; "when in a library, you must not talk on the phone"; "when in a library, you may briefly whisper."

Humans living in social communities function more effectively and peacefully when their actions are guided by a shared set of norms [Bic06, UM77]. The ability to represent and follow norms has many advantages: Norm-consistent actions increase multi-party coordination and cooperation and thus benefit the community

as a whole. Norms also simplify people's action selection and standardize behaviors across time and generations. And norm-consistent actions are more predictable and understandable [MSA17b].

But how does the human mind represent norms, and how are they activated and learned? Surprisingly, there are few cognitive science approaches to the central phenomenon of norms. Logical and specifically deontological approaches have been proposed to formally represent a system of norms [BAB06, SM14, PS09, Bel10]. These are important starting points, but their formalizations do not necessarily correspond to how norms are represented in the human mind. By contrast, a cognitive science approach would aim at an account of how norms are cognitively represented, how they are activated in relevant situations, and how they are learned in the first place. Here we take a first step toward such an account, following a recent theoretical proposal [MSA17b]. We introduce a basic formal representation of norms that allows us to examine the mentioned cognitive properties of norms (representation, activation, and learning), and we ask what computational models can capture these properties, and what algorithms could learn norms.

Our chapter has three main parts. In the first, we present a novel belief-theoretic norm representation format that explicitly captures the context-specificity of norms and incorporates uncertainty associated with norm representations, using Dempster-Shafer Theory [Sha76]. In the second part, we introduce experimental data on human norm representation and activation that underscore the context-specificity of norms and community members' strong but imperfect agreement (uncertainty) over norm applications. In the third part we use our formal norm representation to ask how such imperfect norms systems can be learned by a computational algorithm that honors several of the critical features of norms, including their context specificity and uncertainty.

6.2 A Representation Format for Norms

We begin by briefly outlining our norm representation format in first-order logic and provide some intuitions as to how context and uncertainty are accounted for in the format. The purpose is to introduce some terminology and a minimal degree of formalism in the proposed approach, which will later be useful in developing an algorithm that can learn norms.

Consider a first-order alphabet $\mathcal{L}$, in which we have all the standard symbols (variables, predicates, functors) and logical connectives. In a deontic alphabet, we further include $\mathbb{O}, \mathbb{F}, \mathbb{P}$ that denote modal operators (generally, $\mathbb{D}$) for obligatory, forbidden and permissible, respectively. In this alphabet, we define a norm, as follows:

Definition (Norm). A norm is an expression of the form:

$$\mathcal{N} := C_1, \ldots, C_n \implies (\neg)\mathbb{D}(A_1, \ldots, A_m),$$

where C represents context conditions and A represents actions or states. The norm expression states that when the contextual atoms C_i are true then the Actions or States A_j are either obligatory, forbidden or permissible, or their negation.

This type of norm definition follows an approach to normative reasoning and norm formalism that some of us have taken previously [MSA17b, BAB06, SM14].

In this chapter, we expand the above representation format by explicitly accounting for uncertainty of a norm as follows:

Definition (Belief-Theoretic Norm). A belief-theoretic norm is an expression of the form:

$$\mathcal{N} := [\alpha, \beta] :: C_1, \ldots, C_n \implies (\neg)\mathbb{D}(A_1, \ldots, A_m),$$

where $[\alpha, \beta]$ represents a Dempster-Shafer uncertainty interval, with $0 \leq \alpha \leq \beta \leq 1$.

Example Consider an example of an agent reasoning about actions it can perform or states it can enter in a library. We can represent this scenario as a Belief-Theoretic Norm System, $\mathcal{T}$, as follows:

$\mathcal{N}_1 := [0.9, 1] :: in(library, X) \implies \mathbb{O}\ state(X, quiet)$

$\mathcal{N}_2 := [0.8, 0.95] :: in(library, X) \implies \mathbb{P}\ action(X, reading)$

$\mathcal{N}_3 := [0.9, 1] :: in(library, X) \implies \mathbb{F}\ action(X, yelling)$

$\mathcal{N}_4 := [0, 0.3] :: in(library, X) \implies \mathbb{O}\ action(X, talking)$

$\mathcal{N}_5 := [0.3, 0.6] :: in(library, X) \implies \mathbb{F}\ action(X, talking)$

The norms in this example have intuitive semantics. They generally state that when agent X is in the library (i.e., $in(library, X)$), then the norm is activated and the agent is obligated to enter a certain state (e.g., $state(X, quiet)$) or prohibited from performing a certain action (e.g., $action(X, talking)$. The location of the center of the uncertainty interval generally suggests the degree of truth of the norm applying and the width of the interval generally suggests the level of support or evidence for that norm. So norms $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ have tight uncertainty intervals close to 1 indicating a confident support for their truth. Norm $\mathcal{N}_4$ states that the action of "talking" is obligatory in libraries. Although the uncertainty interval for this norm is tight, the center is closer to zero indicating confident support for the falsity of the norm. Finally, in rule $\mathcal{N}_5$ the question of whether talking is forbidden in a library may be more uncertain, generating a wider interval centered close to 0.5, indicating support for both truth and falsity, but a general lack of confidence in the evidence.[1]

A belief-theoretic norm system of this form allows the separation of evidence from the norms themselves. The evidence may come in different forms across different modalities and from different sources. The norm system, however, displays the agent's current level of belief about a set of norms that are influenced by the evidence.

[1]The use of deontic logic for normative reasoning is the subject of active debate. Although further discussion of this debate is outside the scope of this chapter, we note that our proposed approach does not require using deontic operators. We can still reason about norms and learn them using the schema described in Definitions 2 and 3. We would simply need to replace the deontic operators and modify the predicates slightly. Norm $\mathcal{N}_5$ in example 1 would become:
$\mathcal{N}_5 := [0.3, 0.6] :: in(library, X) \implies forbidden(X, talking)$

In any given situation, the agent may not be reasoning with every norm in a norm system. Instead, the agent may consider a subset of the system, perhaps including only norms that are applicable to the current situation. We capture this intuition in a *norm frame*, defined below.

Definition (Norm Frame). A norm frame $\mathcal{N}_k^\Theta$ is a set of k norms, $k > 0$, in which every norm has the same set of context predicates and corresponds to the same deontic operator. Thus, in Example 1, norms $\mathcal{N}_3$ and $\mathcal{N}_5$ would constitute a norm frame.

We define a norm frame in this way because it allows for cognitive modeling in a situated manner—that is, reasoning about behavior relevant to a specific situation. This context-specificity provides a convenient constraint that can help simplify computation and better capture human norm representations, as introduced next.

6.3 Norm Representation and Activation in Human Data

We are currently engaged in an empirical research program that tests a number of novel hypotheses about the cognitive properties of norms [MSA17b]. Here we summarize two experiments that illustrate some of these properties and provide the learning data for the norm learning algorithm we introduce in Part 3. In the first experiment, participants *generated* norms relevant to a variety of contexts; in the second experiment, participants *detected* norms relevant to those contexts.

6.3.1 Methodology

In the *generation* experiment [KAAM16], participants ($n = 100$ recruited from Amazon Mechanical Turk, AMT) inspected four pictures, one at a time, that depicted an everyday scene (e.g., library, jogging path; see Figure 1 for examples). While inspecting each picture, they had 60 seconds to type as many actions as came to mind that one is "allowed" to perform in this scene (Permissions), or is "not allowed" to perform (Prohibitions), or is "supposed" to perform (Prescriptions). This

between-subjects manipulation of norm type was constant across pictures so that each participant answered the same question (e.g., “What are you permitted to do here?") for all four pictures they encountered.

Figure 6.1: Four sample scene pictures used to elicit norms

To increase generalizability at the stimulus level, the total number of scenes used in the experiment was in fact eight, four that previous participants had tended to describe as locations (e.g., library, cave), and four that they had tended to describe as activities (e.g., jogging outdoors, serving in a restaurant). Each participant was randomly assigned to receive either the “location" set or the “activity" set. Item set made no difference in the results.

The resulting verbal responses were lightly cleaned for spelling and grammatical errors and responses identical in meaning were assigned the same response code, using a conservative criterion so that variants such as “listening" and “listening to music" were counted as distinct. The resulting data structures were then analyzed for consensus (i.e., how many people generated a given response for a given scene) and context distinctiveness (i.e., whether a response generated for one scene was also

generated in a different scene).

In the *detection* experiment, we presented participants (n = 360 recruited from AMT) with the same pictures, four per participant. Along with each picture, we presented 14 actions (randomly ordered, one at a time) that a person might perform in this context. Any given participant's task was the same for each of their four pictures: to consider the particular scene and judge whether each of the 14 actions is either permitted, or prescribed, or prohibited. This norm type was again a between-subjects manipulation and hence constant across pictures. In addition, to increase generalizability, we used two different formulations for each norm type, summarized in Table 6.1. Formulation made no difference in the results.

Table 6.1: Eliciting Probes for Three Norm Types

Norm Type	Probe formulations
Permission	Are you allowed to do this here? Are you permitted to do this here?
Prohibition	Are you not allowed to do this here? Are you forbidden to do this here?
Prescription	Are you supposed to do this here? Should you do this here?

The 14 actions assigned to a given scene under a given norm type (e.g., Library/permitted) consisted of seven "local" and seven "imported" actions. Local actions were the seven most frequently generated actions for the given scene and norm type in the above *generation* experiment—for example, the seven actions most frequently mentioned to be permitted in the library. Imported actions were comprised of top-seven actions generated for *other* scenes (but under the same norm type). Thus, imported actions were still frequent responses to the same norm probe, but in different contexts.[2] Table 6.2 provides an illustration of this selection process.

[2]We ensured that the imported actions were physically plausible in the given scene/context.

Table 6.2: Origin of Selected Actions for *Library* Scene

Action	Origin
Local, permitted	
reading	from top 7 of Library
studying	from top 7 of Library
sitting	from top 7 of Library
checking out a book	from top 7 of Library
learning	from top 7 of Library
being quiet	from top 7 of Library
using computers	from top 7 of Library
Imported, permitted	
eating	from top 7 of Beach
walking	from top 7 of Cave
listening	from top 7 of Boardroom
filling boxes	from top 7 of Harvesting
washing hands	from top 7 of Public Bathroom
running	from top 7 of Jogging
talking	from top 7 of Restaurant

6.3.2 Experimental Results

We begin by highlighting three findings from the *generation* experiment.[3] First, even though people were entirely unconstrained in their norm-guided actions, they showed a great deal of consensus on the most central norms for each scenario. Table 6.3 displays (in column *Consensus*) the seven most frequently mentioned permission norms in two representative scenarios, *Library* and *Jogging*, with consensus computed as the percentage of participants who mentioned the particular action as permitted in the scenario. (The patterns are consistent across other scenarios.) Second, the most consensual norms are mentioned early on; in other words, what comes to mind first is likely to be a consensual norm. Table 6.3 shows (in column *Position*) the average rank position (1 = first, 2 = second, etc.) at which each action was generated, whereby the expected position under a random distribution would be 4.2 for Library and 4.6 for Jogging. Third, the norms generated for the eight scenarios showed remarkable context specificity. Not only do the two illustrated scenes have no norm in common

[3]We focus here on permissions. Prescriptions and prohibitions show very similar patterns overall, but prohibitions differ from the other two norm types in interesting ways (e.g., less consensus, slower activation) that will be treated in a separate investigation.

among their top seven, but of the 56 permitted actions that were mentioned in the top-7 in each of the 8 scenes, only 5 appeared in more than one scene.

Table 6.3: Permission Norms for *Library* and *Jogging* Scenes in the Norm Generation Experiment

Library		
Permitted Action	*Consensus*	*Position*
reading	84%	2.1
studying	68%	1.8
sitting	47%	3.1
checking out books	47%	4.4
using computers	32%	5.3
learning	32%	6.0
being quiet	32%	7.5
Jogging		
walking	87%	1.4
running	87%	1.9
jogging	53%	4.8
talking	53%	5.1
listening to music	33%	4.3
biking	27%	4.7
looking at birds	27%	6.2

Two main results stand out from the *detection* experiment. First, people showed very high consensus in affirming the permissibility of the seven local actions for their respective scenes. For both *Library* and *Jogging*, this rate was 99%; and across all scenes, the number was 97.2%. That is, even though some of these local actions were actively generated as "permissible" by only a third or half of previous participants (see Table 6.3), when directly confronted with these actions, people almost uniformly recognized their permissibility. (Moreover, this recognition was fast, taking only about 1100 ms on average.)

Second, participants clearly distinguished between the local and the imported actions, accepting the latter as permissible at a significantly lower rate. For Library, this rate was 43%; for Jogging, it was 75%; and across all scenes, it was 66.1% (all statistical comparisons to local actions $p < .001$, signal detection discrimination parameter $d' = 1.49$). That is, for a given context on average, 34% of presented actions were judged to be *not* permitted even though they were explicitly deemed

permissible in other contexts.

These results suggest that norms can be activated by static photographs, and people show high agreement in explicitly recounting these norms (generation experiment). In a more implicit setup (detection experiment), people are fast and almost unanimous in affirming the most important norms of a given context and differentiate them well from norms originating from a different context. Thus, both explicit and implicit judgments show substantial context sensitivity. If these are some of the properties of human social and moral norms, how can they possibly be learned, by humans and machines?

6.4 Learning Norms

6.4.1 How Do People Learn Norms?

In learning social and moral norms, people deal with multiple different norm types (permissions, prescriptions, prohibitions), using many different learning mechanisms, and taking input from many different sources. Here we focus on the process of learning permission norms from simple observation, using responses from a sample of community members described earlier in the *detection* experiment. Our main goal is to put our proposed computational framework to a test. In the future we will develop further applications (e.g., learning of obligations or learning from instruction)

Consider a person who has never spent time in a library. Upon entering one for the first time, he observes several people reading, studying, and a few whispering. Some sit at computers, one is eating while sitting in an armchair, although there is a sign that says "No food or drink in the library." Our observer also sees several people at the check-out counter, subsequently exiting the library, where another sign says "Don't forget to check out." Briefly, a younger person runs alongside the stacks but then sits down next to an adult.

The number of people performing each behavior, their age, expertise, appearance, perhaps responses from others, and the meaning and force of various physical symbols will all contribute to the speed and confidence with which our protagonist

learns the norms of a library. Below we offer a data representational format that incorporates these and other properties of the norm learning process, a format that can also accommodate partial information and unknown prior probability distributions and that can be extended to other learning mechanisms, such as verbal instruction or trial and error.

6.4.2 Data Representation Format of Norm Learning

Consider a set $S = \{s_1, \ldots, s_n\}$ of n evidence sources. For example, an evidence source s_i could be a student in the library, the librarian, or a sign at the entrance. To simplify, we are interested in learning about a norm frame $\mathcal{N}_k^\Theta$ comprising k norms (out of a larger possible set) that all share the same deontic type (here, permissions) and the same general context precondition (here, library).

Let an endorsement $e_{i,j}$ be the i^{th} data source's endorsement of the j^{th} norm, where $e \in \{0, 1, \epsilon\}$. The value $e_{i,j}$ is a form of truth assignment, indicating whether the source endorses the norm to be true (1), false (0) or unknown (ϵ). For example, an observation that a student is reading can be interpreted as showing that this student endorses the norm $\mathcal{N}_2$ to be true in this context, hence $e_{i,\mathcal{N}_2} = 1$. The set Φ_{s_i} represents a given source's finite set of endorsements within a given norm frame, such that $|\Phi_{s_i}| = k$.

Informally, for a set of norms in a given context and for a particular source, we can learn about that source's endorsement of each norm; if we also assign a weight (e.g., reliability, expertise) to the source, we form a *data instance.* Multiple data instances (i.e., evidence from multiple sources) form a data set. More formally:

Definition (Data Instance). A data instance $d = (\mathcal{N}_k^\Theta, s_i, \Phi_{s_i}, m_{s_i})$ is a tuple comprising a norm frame $\mathcal{N}_k^\Theta$, a specific source s_i, a set of endorsements Φ_{s_i} provided by that source, and a mass assignment m_{s_i} corresponding to the amount of consideration or reliability placed on source s_i.

Definition (Dataset). A dataset $\mathcal{D}$ is a finite set of n data instances $\{d_1, \ldots, d_n\}$.

Some of the desirable properties of the proposed data representation format are that we can accommodate various types of sources (e.g., behavior, verbal responses, signs and symbols), differential source reliability (mass), order effects (updates can be tuned, if necessary, to the order of received data), missing and imprecise information (we use ϵ to represent ignorance), lacking prior probability distributions (we do not require any priors), and varying norm dependencies (e.g., we can capture a correlation between the prohibition to yell and the prohibition to talk).

6.4.3 Algorithmic Learning of Experimental Data

We can now apply this representation format to the *detection* data we introduced earlier. The detection experiment featured, for each scene, a norm frame $\mathcal{N}_k^\Theta$ with $k = 14$ potentially permissible actions, where half of the potential actions had been specifically identified as permitted in this scene and the other half as permitted in other scenes (see Table 6.2). Each participant, s_i, indicated whether each of 14 actions was in fact allowed in this scene, providing responses of yes (1) or no (0) or no response (ϵ), thus forming a set of endorsements Φ_{s_i}, with $|\Phi| = 14$. In this particular case we treat all sources as equally reliable, hence carrying identical m_{s_i} weights.

With these representations in hand we can formally define the *norm learning problem* within our framework and set the stage for an algorithm to analyze evidence and derive a norm structure for a given context in a given community. We remind the reader that, according to Definition 2, any norm (e.g., with respect to reading in a library) has an uncertainty interval $[\alpha_1, \beta_1]$ associated with it, which reflects the quality and consistency of the evidence for a given norm to hold. The learning problem thus becomes a parameter learning problem for discovering the values of the uncertainty interval for each norm in a norm frame:

Definition (Norm Learning Problem). For a norm frame $\mathcal{N}_k^\Theta$ and dataset $\mathcal{D}$, compute the parameters $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$ of that norm frame.

As noted earlier, each data instance d represents a potential arrangement of

true and false values for each of the norms in a frame. Setting aside the possibility that $e_{i,j} = \epsilon$, each data instance thus provides a k-length string of 1s and 0s (a given participant's response string in the detection experiment). This string is a sample of the normative endorsements in the given community. The norm learning algorithm represents each string as a hypothesis in a set of hypotheses (termed Frame of Discernment in Dempster-Shafer theory) and assigns uncertainty parameters to each norm, updating those values as it considers each new data instance. **Algorithm 1**, displayed below, achieves this form of norm learning from a human dataset.

Algorithm 6.1 getParameters($\mathcal{D}, \mathcal{N}_k^\Theta$)

1: $\mathcal{D} = \{d_1, \ldots, d_n\}$: Dataset containing n data instances for a norm frame
2: $\mathcal{N}_k^\Theta$: An unspecified norm frame containing k norms $\mathcal{N}$
3: Initialize DS Frame $\Theta = \{\theta_1, \ldots, \theta_{2^k}\}$
4: $m(\Theta) = 1$
5: **for all** $d \in \mathcal{D}$ **do**
6: **for all** $\mathcal{N} \in \mathcal{N}_k^\Theta$ **do**
7: Set learning parameters p_1 and p_2
8: $Bel(\mathcal{N}|d) = \frac{Bel(\mathcal{N} \cap d)}{Bel(\mathcal{N} \cap d) + Pl(d \backslash \mathcal{N})}$
9: $Pl(\mathcal{N}|d) = \frac{Pl(\mathcal{N} \cap d)}{Pl(\mathcal{N} \cap d) + Bel(d \backslash \mathcal{N})}$
10: $Bel(\mathcal{N})_{new} = p_1 \cdot Bel(\mathcal{N})_{prev} + p_2 \cdot Bel(\mathcal{N}|d)$
11: $Pl(\mathcal{N})_{new} = p_1 \cdot Pl(\mathcal{N})_{prev} + p_2 \cdot Pl(\mathcal{N}|d)$
12: **end for**
13: Set frame Θ with $Bel(\mathcal{N})_{new}$ and $Pl(\mathcal{N})_{new}$
14: **end for**
15: **for all** $\mathcal{N} \in \mathcal{N}_k^\Theta$ **do**
16: $\alpha_\mathcal{N} \leftarrow Bel(\mathcal{N})$
17: $\beta_\mathcal{N} \leftarrow Pl(\mathcal{N})$
18: **end for**
19: **return** $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$

The algorithm iterates though each data instance in the data set (line 5) and, per instance, through each norm in the norm frame (line 6). For each iteration, we first set the hyper-parameters p_1 and p_2 (line 8) that specify how much weight the algorithm will place on previous learned knowledge (p_1) and on each new data instance (p_2). These hyper-parameters are then used to compute a conditional belief and plausibility for a norm given that particular instance of data (lines 9,10). The conditional beliefs and probabilities then yield an updated belief and plausibility for each norm (lines 11, 12). Finally, the algorithm updates the uncertainty interval for each norm with the new belief and plausibility values.

The result is a set of belief-theoretic norms (norms accompanied with uncertainty intervals), where the width of the uncertainty interval indicates the amount of support for the norm (which may vary, for example, as a function of number of respondents in the human data sample) and the center position of the interval should correspond to the level of agreement in the human respondents' endorsement of the norm.

To put this algorithm to the test, we selected, from our detection experiment, a norm frame of 6 (out of 14) actions for the context of *Library* and a frame of 6 (out of 14) actions for the context of *Jogging Path*. However, we wanted to capture the context specificity of norms and constructed the frames such that 4 actions (running, sitting, walking, and washing hands) were the same in each frame, albeit differentially endorsed in the two contexts (e.g., running was clearly not permissible in *Library* but very much permissible in *Jogging*). Thus, the algorithm had to track the norm value of a given action not in general, but conditional on the specific context. If the algorithm succeeds it should recognize which actions people consider permissible and which ones they consider impermissible, for each of the two contexts, and even for those actions that occur in both contexts.

Figure 6.3 illustrates this success. We display single runs of the algorithm across the dataset. In the single runs, the algorithm considers each data instance (each of 30 participants' judgments) in each context once (in a fixed order), leading to wide uncertainty intervals at first, but narrower ones as the number of data instances increases (up to the maximum of 30) . We also performed iterative runs (not shown), in which the algorithm considers the dataset multiple times, each time randomly selecting a possible order of instances, and converging on an optimal estimate of the norm endorsements in the given community. These estimates are highly comparable to the end points of single runs after 30 data instances.

6.5 Context-Shifting

In this next section, we extend the ideas stated thus far to consider how agents can learn norms when there is contextual uncertainty. Specifically, we propose a norm representation scheme (shared by all agents) that introduces a novel deontic modal operator, which equips deontic logic with *context-specificity* and *uncertainty* by relying on a formal framework called Dempster-Shafer theory [Sha76]. We also provide an algorithm that learns norms and honors several critical properties of human norms. Crucially, we suggest that the proposed approach provides the necessary computational infrastructure needed for agents to learn a set of shared norms, as well as to communicate about and address individual differences.

6.6 Context-Specific, Belief-Theoretic Norm Representation

The cognitive science literature offers few investigations into the structure and representations of norms. Logical and deontological approaches have introduced formal representations of norm systems [BAB06, SM14, PS09]. Though such formalizations suggest possible representations, they do not necessarily capture the properties that characterize actual human norm representation and learning.

6.6.1 Mathematical Formulation of a Norm System

We define the logical form of norms as follows:

Definition (Context-Specific Norm). A context-specific norm $\mathcal{N}$ is an expression of the form:

$$\mathcal{N} \stackrel{\text{def}}{=} \mathbb{D}_C A \tag{6.1}$$

for a formal language $\mathcal{L}$ together with deontic modal operators for obligatory ($\mathbb{O}$), forbidden ($\mathbb{F}$) and permissible ($\mathbb{P}$), respectively (collectively, denoted by $\mathbb{D}$). $C \in \mathcal{L}$ represents a context condition, and $A \in \mathcal{L}$ represents an action or state. The

norm expression states that in context C the action or state A is either obligatory, forbidden, or permissible, or not obligatory, forbidden or permissible.

This norm definition builds on an approach to normative reasoning and norm representation that some of us have taken previously [MSA17b, BAB06, SM14, SSA+17]. But we extend this approach by explicitly accounting for uncertainty about a norm representation as follows:

Definition (Belief-Theoretic Norm). A context-specific belief-theoretic norm is an expression of the form:

$$\mathcal{N} \stackrel{\text{def}}{=} \mathbb{D}_C^{[\alpha,\beta]} A \tag{6.2}$$

where $\mathbb{D}_C^{[\alpha,\beta]}$ is an uncertain context-specific deontic operator with $[\alpha, \beta]$ representing a Dempster-Shafer uncertainty interval for the operator, and $0 \leq \alpha \leq \beta \leq 1$. An uncertain deontic operator reduces to a standard deontic operator when there is no uncertainty, i.e., when $[\alpha, \beta] = [1, 1]$

Example Consider an agent deliberating about actions it may or may not perform in a library. This situation can be represented as a Belief-Theoretic Norm System, $\mathcal{T}$, as follows:

$$\begin{aligned}
\mathcal{N}_1 &\stackrel{\text{def}}{=} \mathbb{O}_{Lib}^{[0.9,1]}\ quiet \\
\mathcal{N}_2 &\stackrel{\text{def}}{=} \mathbb{P}_{Lib}^{[0.8,0.95]}\ reading \\
\mathcal{N}_3 &\stackrel{\text{def}}{=} \mathbb{F}_{Lib}^{[0.9,1]}\ yelling \\
\mathcal{N}_4 &\stackrel{\text{def}}{=} \mathbb{O}_{Lib}^{[0,0.3]}\ talking \\
\mathcal{N}_5 &\stackrel{\text{def}}{=} \mathbb{F}_{Lib}^{[0.3,0.6]}\ talking
\end{aligned} \tag{6.3}$$

The norms in this example have intuitive semantics. They specify that in the library (i.e., *Lib*), the agent is obligated ($\mathbb{O}$) to be in a certain state (*quiet*) or is prohibited ($\mathbb{F}$) from performing a certain action (*talking*), each with associated uncertainty intervals. The center of the interval denotes the point estimate of subjective certainty for the believed applicability of the deontic operator, and the width

of the interval denotes the level of evidence for that belief. Norm representations $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ have narrow uncertainty intervals close to 1, indicating strong support for an agent's confident belief (e.g., in the obligation to be quiet in the library). Norm $\mathcal{N}_4$ also has a narrow uncertainty interval but a center close to zero, indicating strong support for the belief that the operator does not apply. Finally, $\mathcal{N}_5$ has a wider interval and is centered near 0.5, indicating little evidence for either the belief in the norm's applicability or the belief in its inapplicability.

A belief-theoretic norm system, as proposed, enables the clean separation of norms from the evidence supporting them. The evidence may arrive in disparate forms from different sources and through a variety of sensors. The norm system then represents the agent's current belief about the set of extant norms, in light of the available evidence.

An agent may have a vast set of norm representations, but in any given situation, the agent is unlikely to reason with every one of those norms. The agent may instead consider only a subset of the entire system, such as those norms that apply to the specific context in which it finds itself. This idea is captured by a *norm frame*, defined below.

Definition (Norm Frame). A norm frame $\mathcal{N}_k^{\Theta}$ is a set of k norms in which every norm has the same set of context conditions and the same deontic operator. Thus, in Example 1, norms $\mathcal{N}_3$ and $\mathcal{N}_5$ would constitute a norm frame.

This definition of a norm frame enables us to model an agent's reasoning about behavior in a context-specific manner. Such context specificity constrains and simplifies computation and better captures properties of norm representations in humans, as introduced next.

6.6.2 Evaluation: Dynamic Context Shifting

We selected two smaller norm frames to evaluate the algorithm: six permission norms each for the contexts of library and boardroom. We constructed norm frames such that four permitted actions (reading, talking, walking, and listening) were the same

in each context but with different endorsement rates (e.g., walking was hardly permissible in the boardroom but permissible in the library). If the algorithm captures the context-specificity of norms it would have to track the norm endorsements for any given action not in general but conditional on the specific context. We also included two permitted actions for which we had data for one context but not for the other: using computers in the library and drinking in the boardroom. This allowed us to evaluate the algorithm's ability to handle ignorance and incomplete information.

The algorithm also had to track how these norms are learned when observations are made in dynamic situations involving changing contexts and general contextual uncertainty. In an ideal learning scenario, the agent would obtain a dataset containing a collection of observations from a single context. The generation experiment described earlier (Section 6.3.1 provided such an ideal dataset. However, learning in the real world is far less perfect; data are often obtained in a streaming, unfolding manner through a series of observations made during a certain time window. Observations are made in context, the identity of which might be uncertain and may even change over time.

Consider the example of a norm-learning agent moving through a library between the reception, stacks, and through various conference rooms. At any given moment, the agent may not be entirely certain if it is in one context or another. As it approaches a boardroom, near the threshold, the agent may not be sure if it is within the confines of the boardroom context or within the confines of the general library context. Normative behavior is often learned in this messy manner through observations in changing and uncertain contexts.

Moreover, this contextual uncertainty can influence the agent's normative beliefs themselves, which in turn can result in deviations in normative behavior. Two identical agents who observe the same sequence of actions but differ in their reliability of identifying the context they are in will acquire a different set of norms. The algorithm must be able to track these variations in natural learning and should account for differences in normative beliefs between different agents.

For our experiment, we consider two agents (Agent 1 and Agent 2) that learn

norms by observing actions in a library and a boardroom. For simplicity, we stipulate that at any given moment an agent can either be in the general library context or in a boardroom context. Also, we stipulate that the agents are moving from the general library area into a boardroom. Thus, they first observe actions in a library and then actions in a boardroom. However, the agents differ in their ability to accurately detect the context they are in.

Agent 1 has a reliable context detector and can, with complete certainty, identify its current context. Agent 2 has a more unreliable context detector, at least at the threshold between the general library area and a boardroom. Thus, Agent 2 is initially certain that it is in the library, but as it moves toward the boardroom it becomes uncertain about what context it occupies. Once it is completely in the boardroom it is again certain of its context.

We expect that the learning algorithm can capture not only the context dependency of the norms but also the agents' different normative beliefs (as a result of their different context detection abilities). We further expect that an agent that is more unsure of its context is also less certain about the applicability (truth or falsity) of a given norm. Moreover, we expect that because the agent is unsure of its context at the library-boardroom threshold, it will attribute observations during this time to both contexts (albeit to different degrees), thereby generally acquiring more data instances for norms in each context. This would then have the effect of strengthening the agent's belief in the norm that it learns, allowing for a narrowing of the uncertainty interval.

Figure 1 illustrates the success of the learning algorithm in both capturing the context sensitivity of norms and the differential effect of learning norms in different situations. We display four plots each showing single runs of the algorithm for whether the action of "talking" is permissible. The results show a wider uncertainty interval at first, but narrowing with accumulated data instances, which is consistent with a hypothetical agent roaming the library and then entering a boardroom. We also performed these single runs for the remaining 5 actions over both contexts for both agents. The results at the end of these runs are shown in Table 6.4. Again,

the uncertainty intervals $[\alpha, \beta]$ represent the degree of support for the rule in the observations (α) and the total belief that does not contradict the rule (β).

As predicted, Agent 1's learning of the library norm (Figure 6.3, top left) proceeds by converging to an optimal estimate of the community's norm endorsement but then holds steady once the agent has left the library context and enters the boardroom context. Conversely, this agent's learning of the boardroom norm (bottom left) begins to converge only once the agent has entered the boardroom. The algorithm approaches the descriptive statistics from the experimental data (which it was not given) but maintains a level of uncertainty that reflects the imperfect agreement in the human data.

In the case of Agent 2 learning the library norm (top right), because it is uncertain about the context in the middle of the run (positive sloping part of the red-dotted line), the agent continues to adapt its learning, finally settling on an interval towards the end of the run, which is a later convergence than that of Agent 1. Conversely, the agent's learning of the boardroom norm (bottom right) begins converging earlier than Agent 1. The result of the uncertainty in context is at the end of the run, where Agent 2 settles on an uncertainty interval that does not include the descriptive mean in the data. As predicted, this deviation suggests that Agent 2's learning at the library-boardroom threshold was influenced by both contexts, thereby increasing both the truth of the norm in the library and the falsity of the norm in the boardroom. Moreover, because threshold data is used in both contexts, the number of data points considered are increased, providing a tighter interval.

Agents equipped with the proposed learning methodology can not only dynamically learn from observations in evolving environments but can begin addressing mutual differences in their background knowledge and sensory capabilities. For example, an agent (such as Agent 2) can compare its learned uncertainty interval with human consensus statistics; that way it can either correct its beliefs or refine its context detection accuracy by attending to aspects of the environment it previously overlooked. Agents can also directly compare and contrast uncertainty intervals. For example, agents can agree if their confidence intervals overlap, or if the center of

their intervals are close to each other. That said, being able to perform this sort of introspection would additionally require the agents to be aware of their differences, a challenge that is beyond the scope of this chapter but the subject of future work.

6.7 Conclusion

In this chapter, we presented a formal representation of norms using first-order logic and Dempster-Shafer theory. The representation captures the context specificity of norms that our experimental data suggest are strongly present in humans. Using a data representation format that incorporates several properties of human norm representation and learning, we then developed a novel algorithm for automatically learning context-sensitive norms from the human data. Because the data format is highly generalizable, norms could be learned from different types of evidence sources in different contexts, and explicitly captures uncertainty due to variations in the source's reliability and the quality of the evidence. Moreover, we presented how disparate artificial agents can learn norms in the presence of varying degrees of contextual uncertainty. The proposed representation and learning techniques provide a promising platform for studying, computationally, a wide array of cognitive properties of norms.

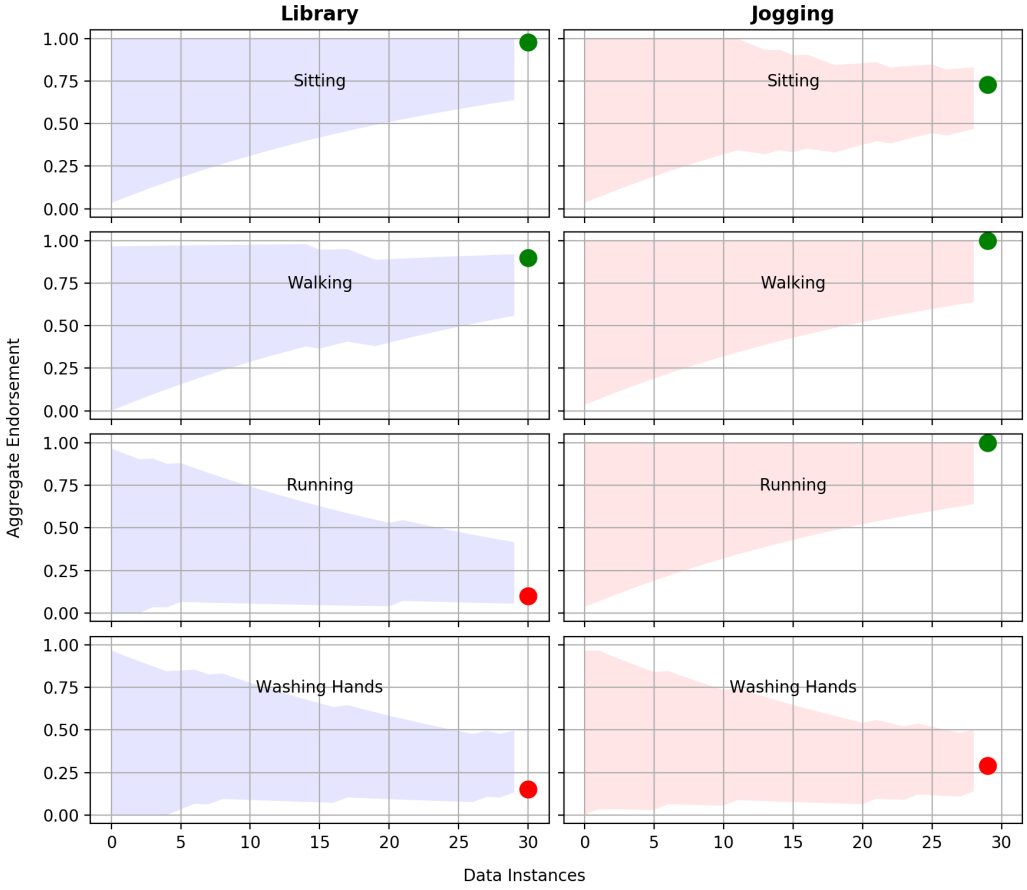

Figure 6.2: Single run of learning across two contexts. The narrowing shaded regions indicate converging uncertainty intervals as new data instances are processed. Filled circles represent the descriptive statistics from the experimental data, indicating actual norm endorsement averages among participants—the proportion of participants who answered yes to the question: “Is this action allowed here?" The algorithm displays convergence towards the descriptive statistics (which it was not given), while maintaining a level of uncertainty reflecting the imperfect agreement within the data.

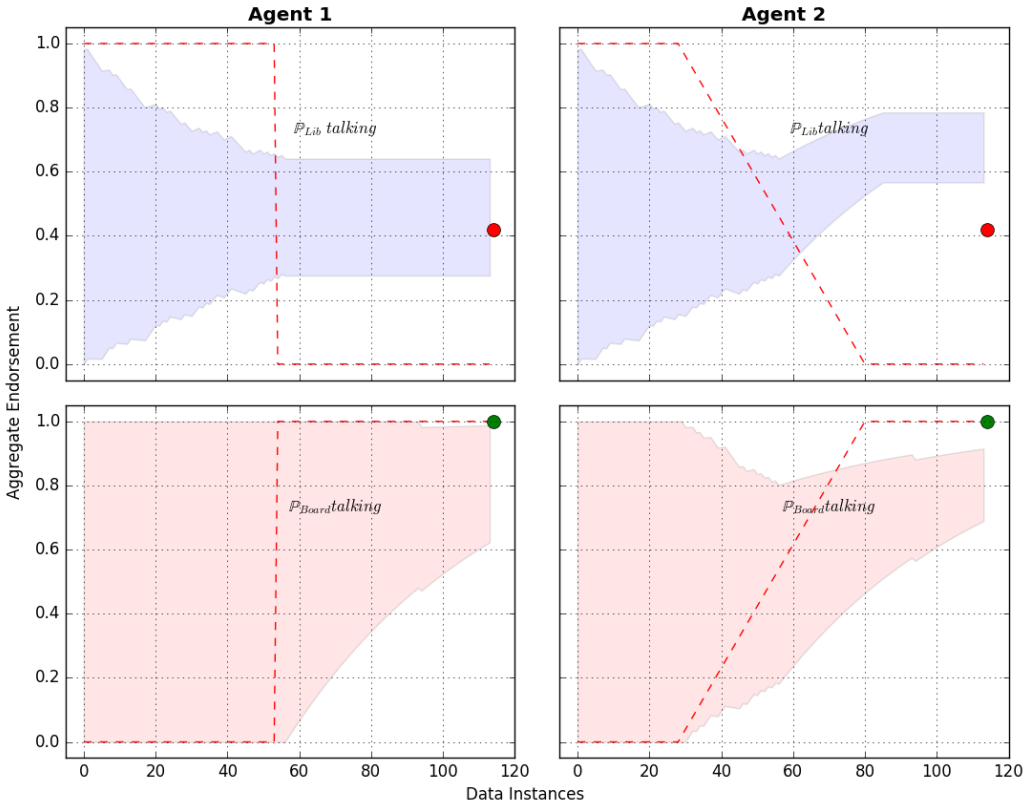

Figure 6.3: Single runs of two Agents moving from a general library context towards a boardroom context. The red-dotted line shows the agent's certainty about its context as it moves. The narrowing shaded regions indicate converging uncertainty intervals as new data instances are processed. Colored dots represent the mean norm endorsements by experimental participants —the proportion of participants who answered yes to the question: "Is this action permitted here?" The algorithm displays flexibility in learning normative rules from observation for various agents with different detection capabilities.

Table 6.4: Uncertainty Intervals for Agents 1 and 2.

Norm	**Consensus**	**Agent 1**	**Agent 2**
$\mathbb{P}^{[\alpha,\beta]}_{Lib}\ reading$	1.0	$[0.63, 1.0]$	$[0.75, 0.97]$
$\mathbb{P}^{[\alpha,\beta]}_{Board}\ reading$	0.93	$[0.59, 0.95]$	$[0.73, 0.95]$
$\mathbb{P}^{[\alpha,\beta]}_{Lib}\ listening$	0.96	$[0.61, 0.98]$	$[0.77, 0.99]$
$\mathbb{P}^{[\alpha,\beta]}_{Board}\ listening$	1.0	$[0.63, 1.0]$	$[0.76, 0.99]$
$\mathbb{P}^{[\alpha,\beta]}_{Lib}\ talking$	0.42	$[0.27, 0.64]$	$[0.57, 0.78]$
$\mathbb{P}^{[\alpha,\beta]}_{Board}\ talking$	1.0	$[0.62, 1.0]$	$[0.69, 0.91]$
$\mathbb{P}^{[\alpha,\beta]}_{Lib}\ walking$	0.91	$[0.57, 0.94]$	$[0.45, 0.67]$
$\mathbb{P}^{[\alpha,\beta]}_{Board}\ walking$	0.32	$[0.21, 0.57]$	$[0.33, 0.55]$
$\mathbb{P}^{[\alpha,\beta]}_{Lib}\ usingComputers$	0.95	$[0.60, 0.96]$	$[0.36, 0.98]$
$\mathbb{P}^{[\alpha,\beta]}_{Board}\ usingComputers$	-	$[0, 1]$	$[0.13, 0.99]$
$\mathbb{P}^{[\alpha,\beta]}_{Lib}\ drinking$	-	$[0, 1]$	$[0.29, 0.89]$
$\mathbb{P}^{[\alpha,\beta]}_{Board}\ drinking$	0.72	$[0.44, 0.81]$	$[0.44, 0.81]$

Chapter 7

Scaling up Norms

In the previous chapter, we discussed how norms can be represented and how agents can learn these norms by observing norm-compliant behavior in the presence of uncertainty. In this chapter, we redefine the formalism and introduce algorithms for learning norms at scale. While retaining the essence of what it means to represent a context-sensitive norm, this chapter dives deeper into the computational aspects of norm learning.

7.1 Introduction

A norm is an instruction to perform (or not perform) a behavior. It has two conditions: (i) it is followed by a sufficient number of individuals in a community (prevalence condition) and (ii) it is one that these individuals expect and demand each other to follow (demand condition or expectation) [Bic06, BdCPD+13, MSA17b]. Norms play a central role in maintaining social order and facilitating coordination and cooperation, without which no human community can exist. As artificial agents become increasingly immersed into human societies, it is crucial that they acquire our existing human norms.

Norms, however, are dynamic and change over time as populations and expectations change, with old norms disappearing and new ones emerging. So, agents cannot be preprogrammed with any particular set of norms prior to entering the

society. Instead, they will need to learn them automatically.

Automated norm acquisition is an active area of research spanning decades and tackling different interrelated questions. One line of questioning asks how norms come about as the population evolves (*norm emergence*) and is interested in the evolution of cooperative behavior - what sorts of behavioral choices and interactions lead to social norms? [ST95, SA07, AH81] A second line of questioning asks how extant norms can be learned (*norm learning/identification*) by a new agent? [ACC+07, OM13, CMOS16]. In this chapter, we focus on this second line of questions with a goal towards designing artificial agents capable of learning existing and established human norms and customs through observation and interaction.

Computational approaches to learning norms include learning normative value functions based on observing other agents' behaviors in an Inverse Reinforcement Learning setting [RDT15], exploring the space of behaviors and deriving an optimal policy using Reinforcement Learning [AML16, KAS18], and inferring norms based on others' plans and goals [OM13, CMOS16]. To infer that a certain observed behavior performed by some agent is a norm (i.e., an obligation or prohibition) requires the learner to consider the possibility that observed agent was non-compliant, that the observation itself was ambiguous or unreliable, or that despite its prevalence, society might not demand it. For example, a behavior might be deemed to be optimal by a large majority of the population and therefore performed by most. This does not mean that it is obligatory, and non-performance of such a behavior, while suboptimal, might not be penalized. Current approaches do not consider these various factors, which we consider to be crucial to norm learning.

In this chapter, we propose an approach based on Dempster-Shafer theory that tracks the agent's uncertain normative beliefs about observed behaviors over time. DS theory allows for more naturally modeling incomplete, ambiguous, and unreliable data. We make three *contributions*: (1) a DS-Theoretic frame representation for normative behaviors (Section 7.2) with a set of algorithms (Section 7.3) operating on this representation for extracting both the prevalence condition and demand condition associated with norms, (2) experimental results in an agent-based

simulation environment to establish performance and computational complexity of the proposed model (Sections 7.4.2 and 7.4.3), and (3) experimental results showing that the DS-Theoretic agent can learn norms more effectively than a Bayesian agent when agents are making observations at large distances or in crowded spaces (Sections 7.4.4 and 7.4.5), situations fraught with perceptual ambiguity and unreliability.

7.2 Representing Norms with DS-Theory

Norms are generally considered to be prevalent expectations of behavior [Hab15]. They can be thought of as a prescribed guide for behavior that is generally complied with by members of a society [Mar77]. Norms concerning a specific behavior exist when the socially defined right to control the action is held not by the actor but by others [CC94]. These definitions collectively point to the idea that instructions – prescriptions and prohibitions (i.e norms) – become associated with behaviors when (i) a sufficient number of individuals in the society follow the instructions, and (ii) there is an expectation or demand placed by others for following the instructions [MSA17a].

There is no general agreement amongst researchers, however, about the most suitable representation format for norms. The choices vary widely from explicit logics [ACC+07, MLSRA+14, CMOS16, SSA+17, MSA17b] to implicit policies or reward functions within a state-transition system [ST95, SA07, BAS15, YLS+16]. We build on the choice of explicit norm representations suggested in Chapter 6, that closely tracks human norm representations. Explicit representations allow an agent to separate learning from decision-making. Therefore, agents can choose to comply or not with a norm when they need to act, allowing for an increased level of autonomy. In Chapter 6, I showed that norms associated with behaviors are activated in contexts, and not one but multiple behaviors (that can co-occur) might be applicable in a context.

In building a formalism for a *norm*, we must therefore consider these above findings: that it is associated with a *physical behavior*, which in turn is linked to a

context as well as the *prevalence* of the physical behavior in a society and whether or not the society expects the physical behavior to be performed, namely the *demand* placed on it. For example, in the context of greeting someone, there may be several physical behaviors available to the agent like smiling and waving. Whether or not the act of smiling is associated with a norm is then a deontic determination (e.g., obligatory, forbidden, optional) of the particular physical behavior of smiling in the context of greeting, which in turn is based on the extent to which smiling while greeting is present in a society and whether the society expects its citizens to smile, for example by sanctioning those who do not smile. Here, we propose that the agent is uncertain about the extent to which smiling is prevalent and the extent to which it is demanded by the society. The agent must observe and update its beliefs about the prevalence and demand associated with smiling, and thereby its belief about whether a norm is associated with the physical behavior of smiling in the context of greeting. A deontic determination of the physical behavior's normative status can only be made once the agent is reasonably confident that its prevalence and demand are above certain thresholds.

More formally, consider a language $\mathcal{L}$, in which we have all the standard symbols (variables, predicates, functions) and logical connectives of first-order logic. Consider an agent that can perform several physical behaviors (ψ). But not all behaviors are applicable in all contexts. So, we can say that a context C contains a set of physical behaviors $\{\psi_1, \psi_2, \ldots\}$. For a particular physical behavior ψ in a particular context C, we can define a normative behavior as follows:

Definition (Normative Behavior).Let the tuple $\mathcal{B} = (\psi, C, \mathcal{P}, \mathcal{D})$ represent a normative behavior, where C and ψ are ground atomic formulas in $\mathcal{L}$ representing context condition (C), and the physical behavior (ψ) possible in that context.[1] $\mathcal{P}$ is the prevalence condition which we represent as an uncertainty interval $[\alpha, \beta]$ capturing the epistemic uncertainty that ψ *occurs* in context C, where α rep-

[1]In Chapter 6, I represented behaviors as logical implication rules. The current representation is intended to be more general, without carrying over the cumbersome and controversial formal semantics of logical implications.

resents the belief and β represents the plausibility as defined under DS-theory. $\mathcal{D} = \{true, false, unknown\}$ represents the demand condition, which we consider here to be either verifiable via observation (true or false) or unknown/unmeasurable. As we will see later in this chapter, one way to operationalize the demand condition is by observing sanctioning of a physical behavior. In such cases, either the agent can verify the occurrence or lack thereof of sanctioning after a behavior has been performed, or the agent cannot verify the occurrence of sanctioning.

We can now define a norm as mapping between the normative behavior – including the underlying physical behavior, the associated context, the prevalence of the physical behavior in the context in society, and the demands (if any) that have been placed on the physical behavior in the context in society – and a deontic determination of the behavior's normative status. That is the mapping answers the question of whether a certain physical behavior like smiling has achieved appropriate levels of prevalence and demand in the context of greeting for a norm of, say it being obligatory.

Definition (Norm). A norm $\mathcal{N}$ is a mapping from a normative behavior $\mathcal{B}$ to a normative (i.e.,deontic) status $\{\mathbb{F}, \mathbb{O}, unknown\}$, defined as

$$\mathcal{N} = \begin{cases} \mathbb{F} & \mathcal{B} = (\psi, C, \mathcal{P} = [\alpha \to 0, \beta \to 0], \mathcal{D} = true) \\ \mathbb{O} & \mathcal{B} = (\psi, C, \mathcal{P} = [\alpha \to 1, \beta \to 1], \mathcal{D} = true) \\ unknown & otherwise \end{cases} \tag{7.1}$$

Per this definition, if the demand condition has been met, and the prevalence nears an uncertainty interval of $[0, 0]$, then the norm returns a value of forbidden $\mathbb{F}$. For example, if physical behavior rarely occurs in society and for those cases where it occured, it was sanctioned, then the normative behavior is likely prohibited. If the demand condition has been met, and the prevalence nears an uncertainty interval of $[1, 1]$, then the norm returns a value of obligatory $\mathbb{O}$. For example, if the physical behavior occurs frequently and for those cases it did not occur, it was sanctioned, then the normative behavior is likely obligatory. The norm cannot determine a

deontic value for a normative behavior if the demand condition cannot be met, and the agent can take that to mean the underlying physical behavior is optional. That is, while high prevalence or low prevalence might suggest the possibility of a norm, the agent cannot be sure that the society actually demands its citizens to perform that behavior. For example, walking on the left on a sidewalk might be what everyone does, but it is not an obligatory norm unless the members of the society place demands on others to comply.

Here, threshold levels for prevalence $\mathcal{P}$ are defined as limits ($\rightarrow$) of the values of α and β approaching 0 or 1. however, threshold values for α and β can easily be specified more precisely as specific numeric values in the interval $[0, 1]$.

We will next explain how an agent can infer if a behavior $\mathcal{B}$ is a norm $\mathcal{N}$ from making several observations of ψ in context C.

7.2.1 Basic Notions in DS-Theory

DS-theory is a measure-theoretic mathematical framework that allows for combining pieces of uncertain evidential information to produce degrees of belief for the various events of interest [Sha76]. In DS-theory a set of elementary events of interest is called a *Frame of Discernment* (FoD). The FoD is a finite set of mutually exclusive events $\Theta = \{\theta_1, ..., \theta_N\}$. The power set of Θ is denoted by $2^\Theta = \{A : A \subseteq \Theta\}$. Each set $A \subseteq \Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping (mass) $m_\Theta(\cdot) : 2^\Theta \rightarrow [0, 1]$ such that $\sum_{A\subseteq\Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set $\mathscr{F}_\Theta$. The triple $\mathcal{E} = (\Theta, \mathscr{F}_\Theta, m_\Theta(\cdot))$ is called the *Body of Evidence* (BoE). For ease of reading, we sometimes omit $\mathscr{F}_\Theta$ when referencing the BoE. Given a BoE $(\Theta, \mathscr{F}_\Theta, m_\Theta(\cdot))$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B\subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) = 1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict

A. The *uncertainty* interval of A is $[Bel(A), Pl(A)]$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$.

7.2.2 Relationship to Bayesian Theory

DS-theory is consistent with classical probability theoretic notions of interval probabilities [FH91], and is often considered a generalization of Bayesian theory.[2] It is the framework of choice for the data fusion community. But, an uncertainty processing framework must be selected based on the needs of the particular application [ST85]. Here, we justify our choice of DS-theory.

Consider the urn problem, in which an urn contains an unknown number of red, yellow and blue balls. We want to predict the color of a ball that we pick out. We learn about this urn by sampling (with replacement) one ball at a time and updating our beliefs. Our beliefs can then provide a quantifiable basis for subsequent predictions.

In a Bayesian framework, we model this example as a probability space $\Omega = \{r, y, b\}$, assign an initial uniform prior over measurable events, and then use Bayes' rule to update it based on our observations. This handles the *randomness* aspect of uncertainty. In DS-theory, we would establish an FoD $\Theta = \{r, y, b\}$ and update masses based on observations and a generalized form of Bayesian conditioning per [FH13], without assuming a uniform prior.

Now, what if we were blue-yellow color blind? We would not be able to distinguish between blue and yellow balls. This is the *ambiguity* aspect of uncertainty. In a Bayesian framework, we could split the weight assignment between the two colors evenly, but this is an additional distributional assumption, over and above assuming a uniform prior, that might be unsupported. A DS-theoretic approach handles this more naturally because, unlike Bayesian theory, it allows for set-valued random variables. A mass can be assigned to the set $\{y, b\}$ without needing to decide how to split the measures.

[2] The close connection with Bayesian theory is one of the reasons we selected DS-Theory over alternatives.

Similarly, what if we used an unreliable sensor to determine the color? This is the *ignorance* aspect of uncertainty. A sensor that predicts blue with 50% reliability means that 50% of the time it predicts accurately, but the rest of time it could be right or wrong. In a Bayesian framework there is no way to assign a weight of 0.5 to true ignorance (the entire probability space) without having to decide how to spread the weights. A DS-theoretic approach handles this more naturally by assigning $m(\{blue\}) = 0.5$ and $m(\Theta) = 0.5$.

To learn norms in the real world we need to account for occluded observations (ambiguity) and variable sensor reliability (ignorance). A DS-theoretic approach handles this more naturally than a Bayesian one; it is therefore our choice.

7.2.3 Bundling Behaviors into Contexts with Indexed FoDs

As noted earlier, each context C has an associated set of physical behaviors relevant to it $\{\psi_1, \psi_2, \ldots\}$[3]. When observing these physical behaviors in context, agents often receive evidence for one or more of these physical behaviors together, as a set. For instance, when an agent is collecting data by observing greetings, it may also observe waving, smiling, talking etc. Together these physical behaviors may be the only ones customarily relevant within the greeting context. We first need to reason about the prevalence of these physical behaviors in the context of greeting. We do so by reasoning about the uncertainty associated with the occurrence of these related physical behaviors together within a DS-theoretic FoD Θ, indexed by context C,

[3]From Chapter 6, we know that normative behaviors are context-specific and are somehow activated by characteristic features of a given context. The property of context sensitivity requires that the agent recognize what context it is in and activate the context appropriate behaviors and norms. It is currently unknown exactly how humans recognize contexts and how context activates the relevant normative behaviors. One way is that the agent classifies the scene as being a particular context, and then activates the normative behavior set that is relevant for that context. During learning, however, the agent does not yet know the normative status of any of its behaviors. It must start somewhere. The agent starts with a set of physical behaviors it can perform. For each context, certain subsets of these physical behaviors might be relevant (in theory all physical behaviors could be relevant for a context – while "greeting" you could, in theory, do a backflip). For each context, a set of normative behaviors can be formed: each normative behavior corresponds to a physical behavior in that particular context. A particular physical behavior could be associated with multiple contexts and so it might be linked to multiple normative behaviors (one for each context). Whether or not this normative behavior has a deontic determination is a different question, one that is answered during the learning process by tracking prevalence and demand. Here, we do not discuss this context-sensitivity as we are focusing on learning prevalence and demand, so we are within a single context and we assume that all physical behaviors are relevant in this context.

called an "indexed FoD."

For example, in the context of *greeting*, the subset of applicable or relevant physical behaviors might be $\{\psi_1 = smile, \psi_2 = wave\}$. We incorporate them into an indexed FoD, $\Theta_{greeting}$ comprising all possible combinations of the behavior propositions (and negations) present in the selected subset of behaviors:

$$\Theta_{greeting} = \{(smile, wave), (smile, \neg wave), (\neg smile, wave), (\neg smile, \neg wave)\}$$

Modeling the FoD in this exhaustive way ensures that elementary events in the frame are mutually exclusive of each other. Now, if we are interested in measuring the amount of support in favor of, say behavior $\mathcal{B}_2$[4], based on our evidence that is captured in the frame $\Theta_{greeting}$, we would measure $Bel(\{(smile, wave), (\neg smile, wave)\})$ as this captures the level of support for just *wave* irrespective of *smile*, from a body of evidence that captures both.

More generally, a DS-theoretic elementary event θ, is a tuple of all the behaviors ψ (or their negations) present in the context. A set of elementary events forms an indexed frame of discernment Θ_C.

Definition (Indexed FoD). Consider an index $I = \{1, \ldots, k\}, k > 0$, where k represents the number of behaviors $\psi_1, \ldots, \psi_k$ in a given context C. Let a set $G \in 2^I$. We define an exhaustive elementary event as a k-tuple $\theta_G = (\tilde{\psi}_1^G, \ldots, \tilde{\psi}_k^G)$, where, for $1 \leq p \leq k$:

$$\tilde{\psi}_p^G = \begin{cases} \psi_p & \text{if } p \in G \\ \neg\psi_p & \text{if } p \notin G \end{cases}$$

An indexed FoD is a set of exhaustive elementary events

$$\Theta_C = \{\theta_G \mid \forall G \in 2^I\}$$

[4] $\mathcal{B}_2$ is the normative behavior corresponding to the physical behavior $\psi_2 = wave$ in context $C = greeting$.

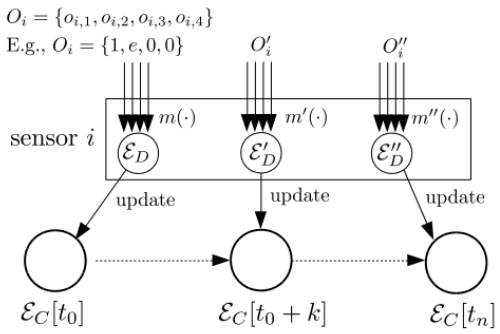

Figure 7.1: Normative agent model. In each time step, new observations arriving at the sensor are packaged as data BoEs, which can then be used to update underlying cognitive BoE. Ambiguity in observation is recorded as ϵ and sensor reliability is tracked with $m(\cdot)$.

7.3 Agent Model of a Norm Learner

We can construct an agent model comprising a set of *cognitive BoEs* each cognitive BoE $\mathcal{E}_C[t] = (\Theta_C, \mathscr{F}_{\mathcal{E}_C}[t], m_{\mathcal{E}_C}(\cdot)[t])$ defined over a corresponding indexed FoD Θ_C. A cognitive BoE represents that agent's knowledge about the uncertainty associated with *a set of* behaviors in a particular context. Initially, the observer is entirely ignorant about the probability of occurrence of any behavior in a particular context. Thus, the cognitive BoE is vacuous, i.e., $\mathscr{F}_{\mathcal{E}_C}[t_0] = \Theta_C$ and $m_{\mathcal{E}_C}(\Theta_C)[t_0] = 1.0$. The observer will need to update this cognitive BoE based on evidence it receives from observations it makes via its sensors. We model this learning and update process as a sensor or data fusion problem. Intuitively, the observer's sensors provide evidence in the form of BoEs, which it can combine with its own cognitive BoE.

7.3.1 Model of Evidence Received

The agent model includes one or more sensors (Figure 7.1). Each sensor is capable of producing a data BoE $\mathcal{E}_D$, which like the cognitive BoE is defined over an indexed FoD, where each $\mathcal{E}_D = (\Theta_C, \mathscr{F}_{\mathcal{E}_D}, m_{\mathcal{E}_D}(\cdot))$[5] is distinct and is not updated. Each sensor i makes a set of k observations $O_i = \{o_{i,1}, \ldots, o_{i,k}\}$ as a form of truth assignment where each $o_{i,j} \in \{0, 1, \epsilon\}$ indicates whether the source determines, for a given

[5]We use $\mathcal{E}_D$ and $\mathcal{E}_C$ as subscripts to masses $m(\cdot)$ and focal sets $\mathscr{F}$ to clearly specify their associated BoE.

context C, whether a certain behavior $\psi_j, 1 \leq j \leq k$ is observed (1), not observed (0) or unknown (ϵ). For instance, an observation that a person in a greeting context is smiling, but not waving can be represented as $O_i = \{1, 0\}$. We can combine the observation O_i with other information about the source as well as a DS-theoretic mass assignment to form a data BoE $\mathcal{E}_D$. Thus, for each set of observations, we can generate a corresponding $\mathcal{E}_D$. For example, an observation $O = \{1, \epsilon, 0, 0\}$ by a sensor with reliability 0.75 is translated into the DS-theoretic proposition $A = \{\{1, \underline{0}, 0, 0\}, \{1, \underline{1}, 0, 0\}\}$. Then, a data BoE $\mathcal{E}_D$ can be created with a focal set $\mathscr{F}_{\mathcal{E}_D} = \{A, \Theta_C\}$ and a normalized mass function $m_{\mathcal{E}_D}(A) = 0.75, m_{\mathcal{E}_D}(\Theta_C) = 0.25$. Thus, through the use ϵ and the mass function, we can capture both ambiguity from say occlusion as well as unreliable sensors, respectively.

7.3.2 Learning Normative Prevalence $[\alpha, \beta]$ through Incremental Update

We employ a conditional update strategy (CUE) that has upgraded Dempster's original rule of combination of evidence to accommodate the inertia of available evidence and address some challenges with respect to conflicting evidence [PMZ+09]. In particular consider a given $A \in \mathscr{F}_{\mathcal{E}_D}$. The updated mass (from iteration t to $t+1$) $m_{\mathcal{E}_C}[t+1] : 2^{\Theta} \rightarrow [0, 1]$ of an arbitrary proposition $B \subseteq \Theta_C$ is:

$$m_{\mathcal{E}_C}(B)[t+1] = \mu[t] \cdot m_{\mathcal{E}_C}(B)[t] + (1 - \mu[t]) \sum_{A \in \mathscr{F}_{\mathcal{E}_D}} \nu(A)[t] \cdot m_{\mathcal{E}_D}(B|A)[t], \ \forall B \subseteq \Theta \tag{7.2}$$

where the CUE parameters $\{\mu[\cdot], \nu(\cdot)[\cdot]\}$ are non-negative real and satisfy

$$1 = \sum_{A \in \mathscr{F}_{\mathcal{E}_D}} \nu(A)[t], \ \forall t, \text{and } 0 \leq \mu[t] \leq 1 \tag{7.3}$$

The conditional in the above equations are derived from Fagin-Halpern conditionals which can be considered an extension of Bayesian conditional notions [FH13]. When $m(B) = Bel(B)$, the conditional mass can be given by:

$$m_{\mathcal{E}_D}(B|A) = \frac{m_{\mathcal{E}_D}(B)}{m_{\mathcal{E}_D}(B) + Pl_{\mathcal{E}_D}(A \setminus B)} \tag{7.4}$$

Beliefs (α) and plausibilities (β) are determined from the Cognitive BoE ($\mathcal{E}_C$). When operationalizing Equation 7.2 within the agent model, we only need to consider and update those propositions B in the data BoE that are subsets of the observed proposition A. This due to the following lemma, the proof for which can be found in [KPD+04]:

Lemma 7.3.1 *Given a BoE $\mathcal{E} = (\Theta, \mathscr{F}, m(\cdot))$ and $Bel(A) > 0$, consider the conditional BBA $m(\cdot \mid A) : 2^\Theta \mapsto [0,1]$. Then $m(B \mid A) = 0$ whenever $\overline{A} \cap B \neq \emptyset$.*

Moreover, we only need to iterate through $\mathscr{F}_{\mathcal{E}_C}$ and $\mathscr{F}_{\mathcal{E}_D}$, and not necessarily all the propositions in 2^{Θ_C}. This is because our data representation of observations O limits the sorts of propositions whose mass will ever be non-zero, extending [KPD+04]. These features allow for significant reductions in the computational complexity of the algorithm as will be shown in the experiments later in this chapter. Algorithms 1 and 2 show how this can be accomplished.

Algorithm 7.1 Pseudo code for agent norm learning

```
C: current context
L: lifetime of learning
E_C ← retrieveCognitiveBoE(C)
D ← ∅
while t < L do
    actors ← chooseActorsToObserve(C)
    behaviors, sanctions ← observe(actors)
    E_D ← createDataBoE(behaviors,sanctions)
    E_C ← update(E_C,E_D)
end while
for all behavior B ∈ E_C do
    bel ← getBelief(B)
    pl ← getPlausibility(B)
    D[B] ← classify(bel,pl)
end for
return D
```

Algorithm 7.2 Pseudo code for CUE-based update

$\mathcal{E}_C$: Cognitive BoE
$\mathcal{E}_D$: Data BoE
μ: CUE parameter representing the learning rate
for all propositions $B \in \mathscr{F}_{\mathcal{E}_D}$ **do**
  $sum \leftarrow 0$
  for all propositions $A \in \mathscr{F}_{\mathcal{E}_D}$ **do**
    if $B \subseteq A$ **then**
      $m(B|A) \leftarrow \frac{m(B)}{m(B)+Pl(A \setminus B)}$
    else
      $m(B|A) \leftarrow 0$
    end if
    $\nu \leftarrow m_{\mathcal{E}_D}(A)$
    $sum \leftarrow sum + (m(B|A) \cdot \nu)$
  end for
  $m \leftarrow$ getCurrentMass($\mathcal{E}_C$)
  if $B \in \mathscr{F}_{\mathcal{E}_C}$ **then**
    prior $\leftarrow \mu \cdot m$
  else
    prior $\leftarrow 0$
  end if
  $m_{\mathcal{E}_C}(B) \leftarrow prior + (1 - \mu) \cdot sum$
end for
return $\mathcal{E}_C$

7.3.3 Learning Normative Demands $\mathcal{D}$ from Sanction Signals

Societal demands can be realized through norm enforcement strategies, such as sanctioning. There are a multitude of different mechanisms with which sanctioning can be accomplished including punishments to reduce an agent's fitness or loss of utility or setback in progress towards goal, or even reputation harms and blacklisting, and restricting access to future rewards. Sanctions can also be positive such as reward mechanisms for norm compliance.

Regardless of the mechanism, the goal of sanctioning is to influence the agent's belief about the normativity of a behavior and thereby impact its decision-making with respect to that behavior. We consider those sanction signals (special enforcement actions) that aim to notify the offender that a certain behavior was non-compliant. These sanctions can occur in several different ways. First, the agent is directly sanctioned by another after behavior. Second, the agent observes another agent, and also observes that agent receiving sanctioning. Third, the agent simply observes a behavior and then infers that sanctioning should occur (although it

does not) for non-compliance based on noticing that the behavior was sub-optimal and unlikely to have been chosen by a rational agent had there not been a norm associated with the behavior.

We view each of these three sanctioning processes to be learning experiences for the agent in which the agent follows a general strategy of modifying a behavioral observation captured in the data BoE $\mathcal{E}_D$ associated with the particular behavior. The agent also tracks a value for $\mathcal{D}$, setting it to true if it observes or infers sanctioning, setting it to false if it observes conflicting sanctions and leaving it unknown if there is no sanctioning. For the first type of sanctioning (direct), if the agent experiences, first hand, a sanctioning for a behavior, it produces a new data BoE in which the particular observation $o_{i,j}$ corresponding to the sanctioned behavior is modified to change its value from $1 \rightarrow 0$ if a forbidden behavior was performed, or from $0 \rightarrow 1$ if an obligatory behavior was not performed[6]. The mass function for the collective observation is set to 1.0. For the second type of sanctioning (indirect), the agent modifies the data BoE associated with behavior being observed by changing the value of the observation $o_{i,j}$ as before, but also modifying the mass function based on the strength of the reliability of the reading (for example, depending on how far the agent is from the observed behavior). Finally, the third type of sanctioning (inferred) also modifies the data BoE, however, it changes the observation to ϵ[7], thereby weakening the strength of the sanction signal. Moreover, here too the mass function is modified to capture some of the nuances of reliability.

7.4 Experimental Results

There is no single universally agreed-upon experimental paradigm or baseline for norm learning in multi-agent systems as any given experimental setup is dependent on the choice of the underlying norm representation, which as we have noted varies widely. Evaluating an approach that learns a Linear Temporal Logic (LTL) norm[8]

[6] We assume an agent that is not devious or deceitful.

[7] Recall from Section 7.3.1 that ϵ is assigned when it is unknown if a certain behavior is observed.

[8] An example of an LTL norm might be: "in some present or future time step the agent must perform a particular behavior."

will necessitate the use of state-transition systems, whereas one for a representation such as ours need not do so. Our central claim in this chapter is that the proposed approach can be a useful and feasible way to model different forms of uncertainty during norm learning. We support this claim through a series of agent-based simulations in which we study the dynamics of the belief update process, performance, computational complexity, and the significance of imperfect data.

7.4.1 Simulation Setup

We considered an agent-based model (ABM) simulation environment. The environment consisted of several agents that could observe and act in the environment. We modulated several parameters for each of the experiments and these are shown in Table 7.1. We assumed that at the start of the simulation, the norm system (fixed # *norms*) had reached some stable normative state in the society and certain behaviors had been assigned with a normative status – Forbidden, Obligatory, or Optional, which we assigned randomly. Certain actors (not the norm learners) were aware of the normative status of the behaviors and their choice of performing a particular behavior during a simulation cycle was governed by the normative status of that behavior and the actor's own *compliance rate*. In addition, these agents could sanction each other if they observed non-compliant behavior, governed by their *sanction rate*.

7.4.2 Experiment 1: Dynamics of Belief Update

We first study how beliefs and plausibilities change over the course of a simulation run. Figure 7.2 shows convergence plots for these uncertainty intervals through the simulation under different experimental conditions – compliance, sanction rate and ambiguity. For each of the runs, we considered three behaviors and assigned them a different normative status of forbidden, optional or obligatory.

Where the sanctioning rate is low and compliance rates are high for certain behaviors, a learner might find it difficult to confirm if the behavior is actually a norm. These are instances where the learner might benefit from explicit instruction of normative status, or by inferring a norm is present if the behavior was sub-optimal.

	Experiment 1 (Convergence)	**Experiment 2 (Perf/Complexity)**	**Experiment 3 (Unc − Distance)**	**Experiment 4 (Unc − Occlusions)**
# Actors	100	100	50	30
# Norms	3	[3 - 12]	3	3
Lifetime	10	[25 - 500]	100	100
Compliance Rate	0.1/0.99	0.1/0.99	1.0	1.0
Sanction Rate	0.1/0.99	0.1/0.99	0.99	0.99
Sensor Reliability	0.1/0.99	0.1/0.99	N/A	N/A
Ambiguity Rate	0.1/0.99	0.1/0.99	N/A	N/A
Height/Width	N/A	N/A	250/ 250	250/ 250
Visible Range	N/A	N/A	[10 - 50]	20
Max. Visible Dist.	N/A	N/A	50	50
# Obstacles	N/A	N/A	0	[25 - 150]
# Runs of each setting	1	1000	40	20
Dependent Variable(s)	Uncertainty interval	Prec., Recall, focal set	Iterations	Iterations

Table 7.1: Parameter selections for each experiment in the simulation. The bottom row shows the dependent variable (not modulated) for each experiment.

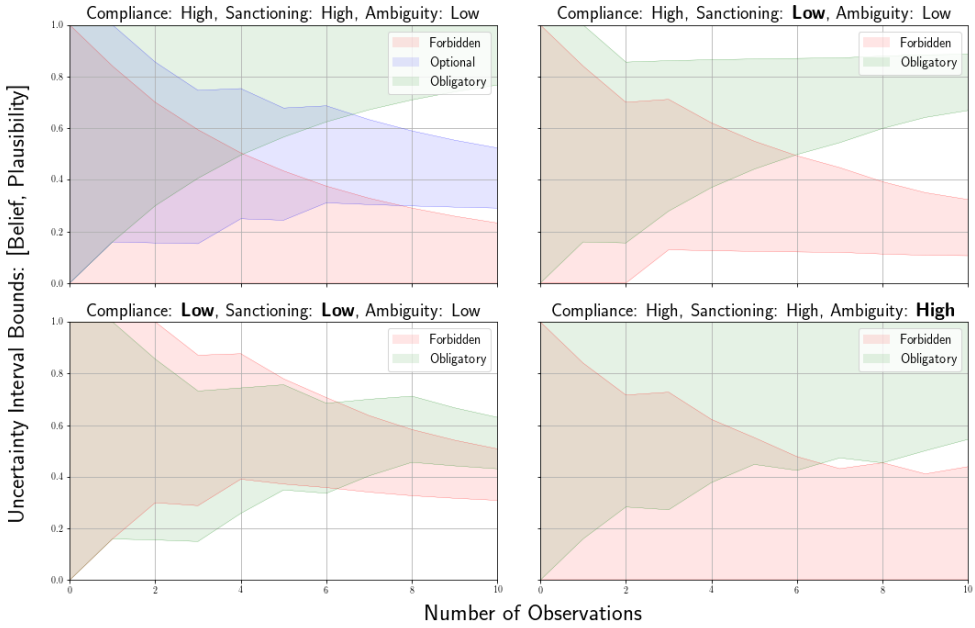

Figure 7.2: Plots showing the ability of the proposed approach to distinguish between qualitatively and quantitatively different types of uncertainty over three types of norms (obligatory, forbidden and optional). We assume complete ignorance at the start ($[0, 1]$). Top left: With low uncertainty, intervals converge to tight bounds according to deontic statuses of three norms, Top right: By lowering the sanctioning rate, intervals move closer to each other (removed optional norm for clarity), Bottom left: By lowering compliance as well intervals overlap and all actions appear optional, Bottom-right: Increasing observational ambiguity widens the interval

In Figure 7.2, the uncertainty that each of the three behaviors occurred is given by the shaded regions. In this case of high compliance, sanction rate and low ambiguity (top left), the uncertainty converges to tight intervals as more observations are made and aligns closely with the pre-established deontic statuses of the three behaviors. For example, the belief and plausibility for the obligatory behavior converged from $[0, 1]$ to much narrower interval of $[0.79, 1]$. The agent was additionally able to declare this behavior as a norm if sanctioning was also observed.

By lowering the sanctioning rate (top right), we observe that while their widths remain largely unchanged, both intervals have moved closer to each other. Fewer noncompliant behaviors were corrected, so obligatory behaviors that were not seen, lowered their beliefs and forbidden behaviors that were seen, raised their beliefs. Moreover, the non-corrected data instance still served as evidence and therefore

affected the plausibility in a similar manner. Because the compliance rates were still quite high, we saw a correct classification despite the fact that the intervals drifted towards the center.

The bottom-left plot shows what happens when the compliance rate drops. The intervals have come together near the center and the observer will not be able to properly classify behaviors as being obligatory or forbidden. In this situation where there many noncompliant actors and reduced norm enforcement due to lower sanctioning rate, essentially all three behaviors appear optional.

If there is ambiguity in the observation (bottom-right), then the intervals will tend to widen, which is consistent with DS-theoretic updates with set-valued random variables, that cannot be directly modeled in Bayesian approaches. Crucially, the widening of the interval suggests a qualitatively and quantitatively different form of uncertainty as that seen in the cases of sanctioning and compliance.

7.4.3 Experiment 2: Performance and Computational Complexity

Performance. We next study the performance of the norm learning as measured by precision and recall of the normative statuses of the behaviors. We performed a larger scale simulation in which we varied lifetimes – 25, 50, 100 and 500, compliance rates (low/high), sanction rates (low/high), ambiguity rates (low/high), sensor reliability rates (low/high) and number of norms – 3, 5, 7, 10, 12. We considered a 1000 random seeds allowing us a total of 320,000 separate runs of each simulation scenario. We calculated the precision and recall for the multi-class classification problem of deciding if a behavior was forbidden, obligatory or optional. Table 7.2 shows the precision and recall for forbidden and obligatory behaviors for various states of uncertainty. We have color coded the table into four quadrants. Our performance exceeds the state of the art Bayesian approaches [CMOS16] (e.g., in those cases that contain low compliance, ambiguity and sanctioning rates, we show average precision and recall near 0.9, significantly higher than those reported earlier).

Table 7.2 shows the precision and recall for forbidden and obligatory behaviors for various states of uncertainty. We have color coded the table into four

					Sanctioning Rate							
					Low				High			
					Ambiguity				**Ambiguity**			
					Low		High		Low		High	
					$\mathbb{F}$	$\mathbb{O}$	$\mathbb{F}$	$\mathbb{O}$	$\mathbb{F}$	$\mathbb{O}$	$\mathbb{F}$	$\mathbb{O}$
Compliance	Low	**Reliability**	Low	Precision	0.44	1.0	0.5	1.0	0.81	1.0	0.94	1.0
				Recall	0.72	0	0.08	0	0.99	1.0	0.59	0.71
			High	Precision	0.43	0.88	0.49	1.0	0.74	1.0	0.91	1.0
				Recall	0.78	0	0.15	0	1.0	1.0	0.78	0.86
	High	**Reliability**	Low	Precision	0.74	0.99	0.91	1.0	0.86	1.0	0.96	0.99
				Recall	0.99	0.82	0.62	0.23	1.0	1.0	0.6	0.73
			High	Precision	0.68	0.99	0.92	1.0	0.79	0.99	0.91	1.0
				Recall	0.99	0.88	0.79	0.37	1.0	1.0	0.78	0.86

Table 7.2: Performance metrics over four forms of uncertainty. Low classification accuracy in the red region as the agent is unsure how to classify behavior in these "wild-west" sort of settings, where all the actions appear optional. Performance in the yellow and green regions were very similar because of the high sanctioning rate, which corrected for any non-compliant behavior. In the blue region, because compliance is quite high, the norm learner cannot know for sure if what it is observing is a statistical regularity or a norm with an expectation - few are being sanctioned.

	Number of Behaviors in Context				
Lifetimes		**3**	**5**	**10**	**12**
	25	3.9	5.4	10.5	12.1
	50	4.8	6.7	14.3	17.8
	100	5.5	8.6	19.1	24.3
	500	6.7	11.6	25.2	33.7

Table 7.3: Average size of focal set $|\mathscr{F}_{\mathcal{E}_C}|$. Although the $|\Theta_C|$ grows exponentially with the number of behaviors, $|\mathscr{F}_{\mathcal{E}_C}|$ (over which our algorithm iterates) only grows linearly with the number of behaviors and sub-linearly with the number of lifetimes as the amount of incoming evidence increases.

quadrants. The red region (like bottom-left Figure 7.2) was where the sanctioning and compliance rates were low. Generally, the classification accuracy is low in the red region as the agent is unsure how to classify behavior in these "wild-west" sort of settings, where all the actions appear optional. Performance in the yellow and green regions were very similar because of the high sanctioning rate, which corrected for any non-compliant behavior. Finally, the blue region corresponds to cases when sanctioning rate was low, but compliance was high. Here, the newcomer observer cannot discern the difference between a norm and a convention. The accuracy rates are still quite high because compliance is high, so the newcomer observer is learning the right norms from observation. However, it cannot discern if the convention it observes is in fact an expected norm.

Computational Complexity. In general, DS-theoretic computations are exponential in $|\Theta|$, and for indexed FoDs they can be doubly exponential in the number of behaviors. We can limit the update complexity in Algorithm 7.2 by only iterating through the focal sets of the cognitive BoE $\mathscr{F}_{\mathcal{E}_C}$ and the data BoE $\mathscr{F}_{\mathcal{E}_D}$. $\mathscr{F}_{\mathcal{E}_D}$ is never greater than two elements (one for the observation and other for Θ_C), and we show in Table 7.3 that $\mathscr{F}_{\mathcal{E}_C}$ does not always grow exponentially. Table 7.3 represents a highly compliant multi-agent system with high rates of sanctioning. If there is an increased amount of conflicting evidence, we expect that $|\mathscr{F}_{\mathcal{E}_C}|$ will increase. We can monitor this rate of growth in $\mathscr{F}_{\mathcal{E}_C}$ and potentially even terminate learning to seek clarification. Generally, when there is substantially more unreliable and ambiguous

data, the computational complexity can grow because the focal set must also keep track of not just singletons but also propositions.

Second, we incorporated a recent recursive approach (called REGAP) to access large FoDs [PPMS16]. However, REGAP requires the maintenance of a mass vector that is of size $2^{|\Theta|}$, which is untenable for our indexed FoD. We modified REGAP to only represent non-zero masses in a hashtable, and adapted the corresponding mass and belief access algorithms accordingly. Even with these complexity reduction measures, the belief and plausibility computations in lines 12-13 of Algorithm 7.1 can still be quite costly for large sets of behaviors [PPMS17]. We are currently exploring ways to reduce the complexity of these operations, as described in Section 7.5.

7.4.4 Experiment 3: Uncertainty due to Learner's Distance from Behavior

Unlike experiments 1 and 2, Experiments 3 and 4 were conducted in a two-dimensional simulation, where the agents were allowed to explore a 2D environment. We were able to factor in the role of distance and occlusions in this simulation environment. In these experiments, we ask if an DS-based agent can learn norms faster than a Bayesian agent. As noted earlier, DS-theory is a generalization of Bayesian theory. We hypothesized that if the DS-agent can use any observation (whether or not entirely reliable) it is more likely to recognize and converge to an interval commensurate with a normative status. Specifically, we consider the following criteria: prohibitions are behaviors whose DS-theoretic plausibility $\beta < 0.25$ and obligations are behaviors whose DS-theoretic belief $\alpha > 0.75$.[9] Experiments 3 and 4 determine how many learning cycles it would take for an agent to converge to an interval that matches this criteria. We modeled a Bayesian agent in this setting as one capable of handling a data BoE $\mathcal{E}_D$ containing observation O, where $o_{i,j} = \{1, 0\}, o_{i,j} \neq \epsilon$ and has a mass function $m(\cdot) = 1.0$ and $|\mathscr{F}_{\mathcal{E}_D}| = 1$. This model of a Bayesian agent is justified as DS-Theory converges to Bayesian theory in the absence of ambiguity

[9] This is consistent with Definition 7.2.

(ϵ) and unreliability ($m(\Theta) > 0$). Even though this model of a Bayesian agent does not capture more sophisticated models (hierarchical Dirichlet models), we chose this naive model as it maps onto current state of the art norm learning approaches (e.g., those used by Cranefield et al.).

In Experiment 3, our independent variable, maximum reliable distance, is the distance up to which the agent is guaranteed to make reliable ($m = 1$) measurements. Beyond this distance and up to a maximum visible distance the DS-agent's mass function was modulated as a function of the inverse squared distance. The Bayesian agent ignored those readings that were not reliable. The distance also impacted the DS-agent's recording of whether or not the observation was unkown. Thus, although the DS-agent could conceivably view more of the environment, its reliability in those readings as well as the reading itself were adversely impacted. The results shown in Fig. 7.3 suggest that despite these adverse impacts, the DS-agent learned significantly faster than the Bayesian-agent in a number of cases, a difference more pronounced when the maximum reliable distance was decreased.

7.4.5 Experiment 4: Uncertainty due to Occlusions in the Learner's Line of Sight

Experiment 4 followed a similar structure to Experiment 3, with the difference that the environment was made significantly more crowded and the maximum reliable visibility was set to a fixed number. While the Bayesian agent ignored the occluded observations entirely, the DS-agent modulated its mass function and observation value based on the extent to which the observation was occluded (i.e., as a function of the angle between the two agents). The results shown in Fig. 7.4 suggest that, as in Experiment 3, here too the DS-agent learned faster than the Bayesian-agent in the presence of obstacles.

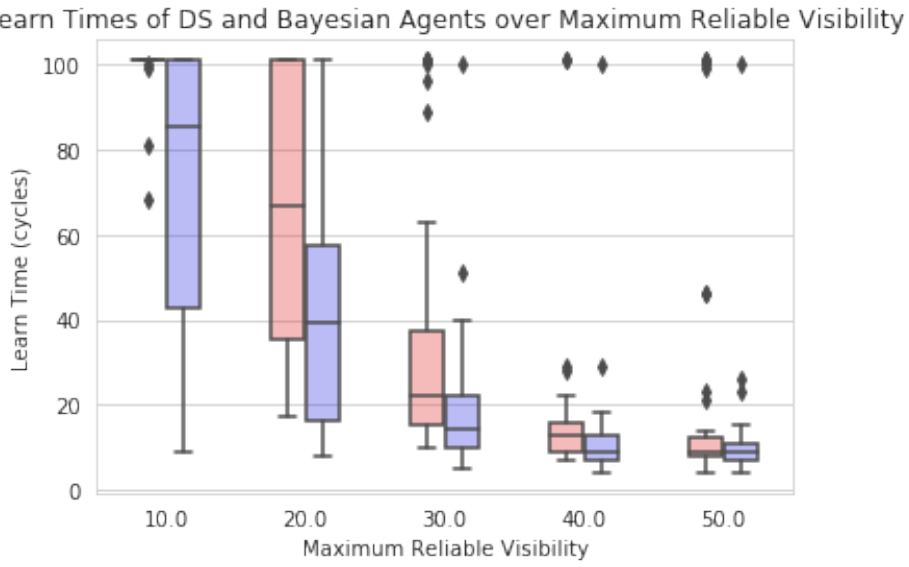

Figure 7.3: Box plot comparing a DS-based agent (blue) with a naive Bayesian-agent (red) when observations are influenced by distance. The results suggest that the DS-agent was significantly faster in learning norms ($p < 0.001$) when the agents were relatively myopic (max reliable visibility < 30). The maximum reliable visibility is the range within which the agents can make observations that are certain. Beyond this range and below a maximum visble distance, the observations are increasingly unreliable and ambiguous.

7.5 General Discussions

In this work, we presented a new approach for norm learning incorporating epistemic uncertainty along with sanctioning during the learning process. The proposed approach provides an explicit representation of normative behavior that can be mapped onto a deontic classification, which, in turn, can be used for norm-compliant decision-making. Experiment 1 provided a close-up view of the dynamics of the norm-learning process, and particularly the convergence behavior of the the agent's uncertainty: beginning with complete ignorance (represented as a $[0, 1]$ interval) and gradually converging as evidence is received. Crucially, the experiment showed that the approach can capture interval uncertainty at any point during the learning process. The true beliefs are always within this interval, a property that is not true of Bayesian confidence intervals. Moreover, the experiment showed how the presence of unreliable data, missing data, and sanctioning can cause the intervals to diverge or shift

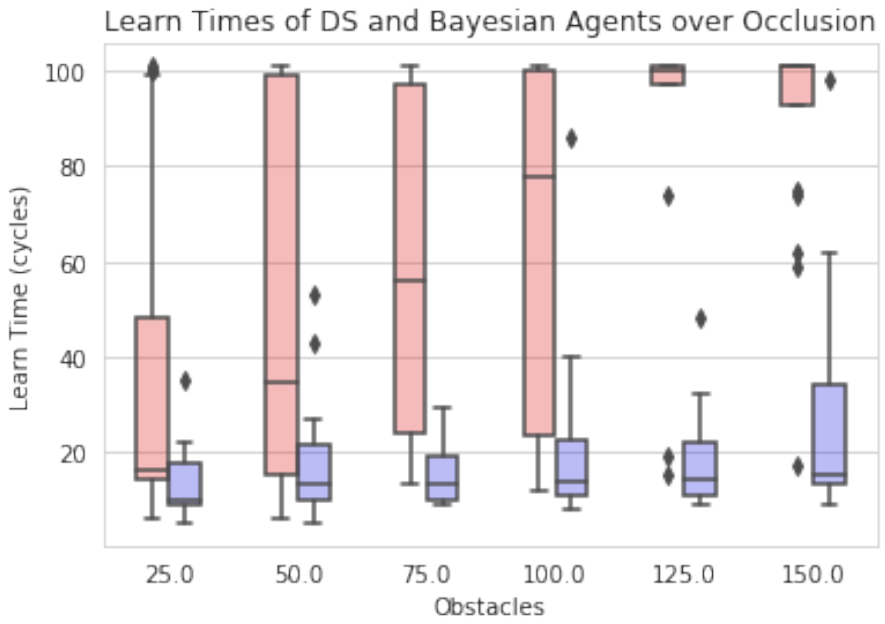

Figure 7.4: Box plot comparing a DS-based agent (blue) with a naive Bayesian-agent (red) when observations are influenced by occlusions in the environment. The results suggest that the DS-agent was significantly faster in learning norms ($p < 0.01$ for 25-75 obstacles, $p < 0.001$ for 100-150 obstacles) when there were obstacles.

in ways that provides insight into the nature of the observations made by the agent. Experiment 2, a larger-scale parameter-sweep, showed how well the approach can correctly classify extant norms and that the approach is computationally tractable, even though traditional DS-theoretic computations are generally not. The approach provides comparable results to existing norm-identification techniques under conditions of uncertainty, while accounting for missing data and unreliable data. We cannot perform direct empirical comparison with other approaches because of differing norm representations. But we can compare the performance of a DS-theoretic approach against a Bayesian one, which we did with Experiments 3 and 4. Experiments 3 and 4 demonstrated the value of not ignoring unreliable and ambiguous data, but instead considering them in a principled way. Specifically, Experiment 3 showed that our approach performs significantly better than a traditional Bayesian approach when agents must make observations at distances, producing unreliable measurements. Experiment 4 showed that the proposed approach performs significantly better than a traditional Bayesian approach when the agent must make

observations in crowded environments, where occlusions can enter the agent's line of sight. Overall, the experiments suggest that the proposed DS-theoretic approach combined with sanctioning is a promising direction for norm-learning research.

7.5.1 Importance of Sanctioning

Our results show that in normative systems where the sanctioning rate is low and compliance rates are high for certain behaviors, a newcomer might be able to correctly perceive these regularities as conventions. However, without sanctioning it is difficult to confirm if the behavior is actually a norm. These are instances where the newcomer agent might benefit from explicit instruction of normative status. Learning norms from explicit instruction via, for example, natural language has been a subject of recent research [WSW20, SOKS18].

We focus on a learning strategy that uses the sanction signals to correct sanctioned observations prior to update. This form of sanctioning has the shortcoming that strong sanctions obtained late in the learning process might not be given the appropriate level of influence to overturn past knowledge. Fortunately, the proposed approach allows for more sophisticated techniques of handling sanctioning like, for example, using the ν parameter in Equation 7.2 to increase or decrease the amount of weight assigned to a new observation or sanction signal, thereby allowing an agent to unlearn bad behaviors quickly.

Sanctioning, while direct, is by no means the only way to confirm the normative status of a behavior. For example, consider inferring that a behavior is a norm when it is not the lowest cost (or most optimal) option that an agent chooses in a particular situation. The shortest path to ice cream may be to walk to the counter of an ice cream truck in a straight line, but the normative one is to choose the higher cost approach, at least from a distance standpoint, and walk towards the end of a long winding line of fellow customers. An agent observing this behavior might be able to infer indirectly the normative nature of this behavior without needing explicit sanctioning. Inferred sanctions are also particularly relevant when there are no other sanction signals like, for example, in a highly compliant or lowly policed

society. In these situations, it is beneficial if a society's normative expectation or demand can still be inferred from the optimality of the behavior. In ongoing work, we are exploring these indirectly inferred forms of norm identification.

7.5.2 Assumptions about the Observations

One advantage of the proposed approach over many other approaches is that the observer makes very few, if any, assumptions about the actors it observes. However, it may be helpful to use some knowledge about the actors' goals or plans in furthering the norm identification process. Building on the notion of indirect norm recognition discussed above, the additional assumptions and knowledge about the actors could assist in establishing the norm status of behaviors in precisely those cases where either sanction signals are unavailable or insufficient.

That said, a limitation of the current work is that the observer knows in advance the set of behaviors applicable in a context. It is able to recognize that certain relevant actions were not performed, or certain performed actions were irrelevant. Lifting this assumption will require the agent to be able to grow or shrink its underlying indexed FoD. Modeling these capabilities under a Bayesian approach would require maintaining a conditional probability table of all the actions in a context. If actions are removed or if new actions are added, these tables have to be recomputed, which is an expensive process. One of the most exciting aspects of the proposed approach is the potential for expanding and growing the set of norms *in real time*, without always having to recompute joint distributions, as would be needed in the Bayesian approach. This is made possible by the notions of "coarsening" and "refinement" provided in DS-theory [Sha76]. Moreover, our approach is strictly "incremental", i.e., we do not need to store the individual observations themselves as we are extending the FoD with new data values; whereas in the Bayesian approach this would be necessary otherwise the distribution cannot be recomputed.

7.5.3 Managing Contexts

One challenge with the proposed approach is the inherent computational complexity of computing various measures of belief and plausibility, especially as the number of behaviors (or norms) increases. Although we propose some ways to address these issues, the complexity of norm identification is still exponential in the number of norms (i.e., set of behaviors within an indexed frame).

One way to circumvent these problems is to leverage a feature of human norm representations, i.e., their context-sensitivity [SSA+17], and construct much smaller indexed FoDs focused on behaviors for narrowly defined contexts. While we focused on a single context setting in this chapter, the approach is already equipped to handle multiple cognitive BoEs allowing an agent to partition its universe of behaviors into smaller and narrower contextual settings. This means that within a given context, we can keep the number of norms fairly small, while allowing the number of contexts to grow as needed. Moreover, a given behavior can exist in several contexts, and contexts themselves could be hierarchically related to other contexts. Sarathy et al. have recently explored norm learning in multi-context settings [SSM17] and we intend to extend the proposed framework as such.

7.5.4 Individual vs System-level Patterns

The bundling of behaviors into contexts also suggests another form of bundling where the observer can aggregate observations with respect to specific actors. In this chapter, the observations are all used to update a single cognitive BoE. However, in some circumstances, it may be helpful to explore the normative status of behaviors at both the agent level and at the society level. Here, cognitive BoEs can be indexed not just by context, but even by the ID of an observed agent. The observer can then track both individual-level behavioral regularities as well as system-level ones. This can help answer questions about whether certain behaviors appear to be conventions or even norms because they are complied with by a small but active group of actors. We can also use these individual differences to explore deviant actors (norms violated

by few) and anomalous norms (behaviors followed by a few, but ardently), where in each case there are significant differences between the statistical regularity of a behavior at the individual level versus that at the system level.

In addition to being able to determine uncertainties associated with specific behaviors, the agent can, with its existing cognitive BoE, answer several additional questions such as whether two behaviors are allowed together? What happens if I perform one of two behaviors and not both? The agent may have never observed these behavioral combinations directly, but it might still have evidentiary support for them indirectly, which is readily accessibly in the cognitive BoE. That said, our representation is largely propositional, thus it has a limited degree of expressivity. This however, does not stop the agent from generalizing and inducing first order atoms from this propositional knowledge. For example, if whispering (which is a specific form talking) is not allowed, then it is likely that talking, generally, is likely not allowed.

7.5.5 Expressivity of Norms

When agents perform actions in the real world, their actions are typically goal-directed, plan-based and change the environment for others. These changed environments, in turn, influence the agent's own decision-making. Our experiments do not show the performance of the proposed approach in these more expressive settings. Past work by Cranefield et al. [CMOS16] and Oren et al. [OM13] demonstrate the strength of plan recognition and representing norms in a more expressive language like LTL to capture these more embodied agents. The focus of our chapter, however, has been along the dimension of improving the quality of uncertainty representations over prior work. In future work, we intend to integrate plan-based approaches with the proposed richer uncertainty model.

7.6 Related Work

Research on normative multi-agent systems (nMAS) is several decades old and extensive [SC11, BdCPD+13]. Since norms play a central role in maintaining social order and facilitating coordination and cooperation, researchers have attempted to formalize and computationalize mechanisms by which norms are formed, propagated and spread through a normative society [SCPP13]. In this chapter, we have focused on norm-formation, and specifically the problem of how individual agents can *learn, identify, or recognize* norms that have already been created in a society, based on observations and interactions [CMOS16, SCPP13].[10] This problem is a particularly important problem once one realizes that it is not enough to simply communicate a set of norms to a new agent prior to entering the society. This is because in dynamic nMAS, agents can enter and leave at any time, norms are constantly changing and evolving, and current agents may not know if the norms they believe to exist, apply throughout the society [HRCS16].

Roughly following the taxonomy proposed by Savarimuthu et al. [SCPP13], there have been two primary past approaches to the norm-identification problem. One approach involves implicitly learning a norm through a series of interactions with current agents in the society using game theoretic and reinforcement learning techniques. Bottom-up learning techniques of this sort were introduced in seminal norm-emergence work by Shoham and Tennenholtz [ST95], subsequently by Sen and Airaui [SA07], and more recently, in work by Beheshti et al. [BAS15] and Yu et al. [YLS+16]. These methods are largely utility maximizing methods, and norms are only represented implicitly (e.g., reward function), so there is no notion of norm expectation. Any strategy that maximizes utility is considered a norm, and agents designed in this approach are not entirely autonomous as there is generally no mechanism by which they can violate norms.

The other class of approaches (including the proposed approach) have focused on modeling norms as cognitive concepts in the agent's mind. These approaches

[10] Note that norm-identification is distinct from the norm-emergence problems of how agents in a society collectively can create norms, or how much do these emerged norms spread.

posit mechanisms that allow agents to recognize, represent, examine, deliberate, and communicate norms, an ability not readily facilitated by the implicit approaches. Andrighetto et al., developed a computational architecture to model norm emergence and identification which allowed for explicit norm representation and learning norms from messages that have been communicated to the agent or those that the agent might have observed [ACC+07, ACCC10, OM13, CACC09]. Andrighetto et al. (2010) and Campenni et al. (2009) allowed for explicit norm representation and learning norms from communication and observation. However, they did not infer obligations, but merely permissions due to the behavioral regularity. Moreover, none of these approaches account for epistemic uncertainty and sanctioning.

There have since been more sophisticated norm representations and formalisms including logical structures afforded by first-order logic, deontic logic, and temporal logics (CTL and LTL) [CANB13, HRCS16, RDF15]. In some approaches, the language of norms is baked into the agent behavior itself where the logical language maps to state-action pairings, allowing for hybrid approach combining this cognitive approach with a reinforcement learning approach [HRCS16].

Certain approaches suggest that norm representations include precondition, action and some deontic operator [MLSRA+14]. Along these lines, some approaches have looked to incorporate plan recognition into nMAS. Work by Oren and Meneguzzi (2014) [OM13], and subsequently by Cranefield et al. (2016) [CMOS16] and Krzisch et al. (2017) [KM17] in Bayesian norm identification have shown success in modeling norm identification and been able to use Bayesian updates to learn norm-compliant behavior from relatively few observations. Oren and Meneguzzi (2013) [OM13] proposed a plan recognition based mechanism, in which their system observes what states the agents always avoid or always achieve. However, this approach assumes that the observed agents' plan libraries are known. Moreover, employing a Bayesian update mechanism requires making assumptions about the prior distribution of norms, which may be unjustified.[11] Cranefield et al. (2016) [CMOS16]

[11]Norm priors need not be uniformly distributed, since they are not derived from a random process.

handled uncertainty in relation to the probability of violation or noncompliance. They use Bayesian updates to learn norm-compliant behavior from relatively few observations. As noted earlier, Bayesian approaches are inherently limited when handling unreliable and ambiguous data.

Work by Sarathy et. al (2017) [SSA+17] (reproduced in previous chapters), Arnold et al., (2016) [AKS17] and Malle et al. (2017, 2019) [MSA17b, MBS19], have suggested that a richer normative system may be needed to account for recent discoveries about human norm representation, activation and learning. In particular, I proposed (in Chapter 6) a computational model that accounts for context and uncertainty. In this chapter, we extended this nascent model by accounting for sanctioning and variable agent compliance in a multi-agent setting. Moreover, we introduce algorithms that significantly improve performance and scalability by reducing complexity for larger sets of behaviors and longer lifetimes.

7.7 Conclusion

Prescriptions and prohibitions (i.e., norms) become associated with behaviors when there is general consensus in society and there is an expectation among members that the norm be complied with. In this chapter, we provide a novel computational approach for an artificial agent to observe behavior and recognize norms associated with these behaviors under uncertainty. The approach accounts for real-world uncertainty where observations are made at different distances, in crowded spaces and where there is variability in consensus and expectations associated with the normative status of behaviors. We show through a series of experiments that the model and approach is state-of-the-art, tractable and provides some insights into how observations can be made in imperfect environments. Doing so, we provide those interested in designing norm-compliant agents with a blueprint for how they might factor in a richer notion of uncertainty when learning norms from observation.

Chapter 8

Consent

Turning our attention away from the cognitive and mathematical aspects of norm representation and learning, in this chapter we focus on recognizing the importance and pervasive nature of one particular type of social norm – consent. "Consent" and its social role in regulating human behavior has been studied heavily in the legal literature. This chapter attempts to make a connection between the centuries-old legal precedent with human-robot interaction, focusing specifically on the types of considerations AI system designers need to account for when building interactive systems.

8.1 Introduction

Social robots are designed to engage people in an interpersonal manner, often as partners, in order to achieve positive outcomes in domains such as education, therapy, and healthcare. Their sociality also aids task-related goals in areas such as coordinated teamwork for manufacturing, search and rescue, domestic chores, and more [BDK16]. HRI research has made tremendous strides in exploring and discovering the social competencies that allow for an "efficient, enjoyable, natural and meaningful relationship" [BDK16]. Understanding what sorts of conduct and behavioral choices enhance or detract from positive social interactions is crucial.

In pursuit of this understanding, HRI research has organized itself along

several lines. There are topics represented as types of interactions (e.g., touch based, long-term), and there are HRI subfields mapped as general aspects of interactions: proxemics [MM17, MM11]; approach [WDWK07]; eye-gaze and joint attention [AS17a]; interruption [THC+16]; touch [VET13, AS18a]; level of politeness [BS16a, BC17, POYI18]. Technical approaches to equipping robotic systems to interact along such lines has generally involved computationalizing an existing social or anthropological theory of interaction applied to the aspect being studied. HRI's empirical research helps evaluate how robots so-equipped interact in experimental settings.

Even as the field has pushed itself to more robust kinds of interactions and more sophisticated systems of planning, the prevalent notion of "social interaction" still focuses largely on dyadic exchanges. These one-on-one engagements usually rely on fairly direct cues and straightforward measurements of user preferences. For example, in approach research, it has been found that people generally prefer robots to approach them from the front and do not prefer robots to approach them from the back [WDWK07]. But there is a major "elephant in the room" in such findings, and in this chapter we propose that it deserves distinct recognition as a way to organize HRI research: the notion of *consent*. Our argument is that HRI's progress in investigating interactive dynamics will not be generalizable enough without tackling the issue of consent head-on.

Consider "approach" again. There are instances when an approach from behind is acceptable and maybe even preferred, for example if a robot was pushing a person out of harm's way from an oncoming car. Wouldn't the human have consented to this act if asked and provided time? Should the robot presume consent? At first glance, it seems that that sort of reasoning is highly circumstantial and potentially too ungainly to be a cogent research problem. Like other subtleties of interaction, it could be chalked up as "context" and tackled at a future date when technology has evolved further. We argue here instead that the tools already exist to directly explore the question further and that understanding consent is central to HRI's future research.

The word "consent" typically brings to mind either notions of legality around highly deviant sexual behavior or around the appropriateness of information given to a patient about a medical procedure. However, consent's rich nuances and intricacies permeate our daily interactions in even normatively neutral environments. For example, when should a robot waiter take your plate at a restaurant? Consent can not only influence a robot's behavioral choice, but also the normative valence of a situation. Consent can turn a normatively neutral situation into a charged one, and vice versa. While consent is one factor in the larger normative fabric of society, it is its normative power and pervasiveness that provides the motivation for HRI researchers to take it seriously.

More generally, social interactions among humans in societies are based on social and moral norms[AS17c]. These norms are deeply ingrained in us and we expect others to abide by them. Now, failing to abide by these norms causes social reactions in humans, from blaming and reprimanding to potentially harsher legal consequences [Sch17]. Given that we have a propensity to anthromorphize robots, especially human-like ones, we are likely to extend these expectations to robots and perceive them as moral agents as well [MS15]. Consent is a key component of this normative setting, continually influencing our expectations of one another. So, it may be natural to expect that ethical competence in robots cannot be achieved without a careful consideration of consent.

Consent is often thought of as a permission (either general or specific) granted to someone to make something happen or perform certain behavior. To consent often means giving this permission, either implicitly or explicitly. While this intuitive definition provides a good starting point, it does not quite capture the intricacies and nuances of consent. How is the permission to be manifested? What does it mean to give this permission? What exactly am I giving permission to? To whom am I giving permission and for how long? Should I be giving permission? What if I would have given permission had I known it was being sought? Would someone else in my shoes have given permission?

As we explicate below, the dimensions of consent can inform and flesh out

existing HRI topics and findings. Returning to the subfields listed above, work on touch has already broached questions of trust, social cues, and the delicate terms on which robots should presume to touch a person [CKTK14, VET13] . Even simple forms of robot-initiated touch evidence gender effects and workplace norms, and are evaluated accordingly [AS18b, WFEŠ16]. Proxemics has broached many consent-related themes, including how people compensate for robots intruding into their personal space [SJWE12]. Spatial preferences, for example how a robot should approach someone, have received sustained attention [SKWD07, KSAO+14], including how people adjust those preferences to aid in a robot's understanding of social signals [MM15]. The gaze of a robot can exert similar force on an interactant's social "space" and shape how a robot's role is perceived [FWL+13, SS17].

Sex robots have directly raised various concerns about consent, including the threat of being "ever-consenting" models of submission to abusive sexuality [Gut12, SA17, SA16]. The risk of neglecting consent as a component of interaction is to risk robots being submissive by default, allowing them to be held up as ideals against people who refuse to consent to abuse. Indeed, some wonder what consent is for such an intimate interaction [FN17]. But it is not just a headline-grabbing issue like sex robots that makes consent salient. The socially assistive role for robots in eldercare residences, especially when those served suffer from dementia, call up numerous questions about "informed consent" and privacy among all stakeholders involved [PNŠ+17, IJVE16]. Consent may thus bind HRI even more strongly to issues for public deliberation around social institutions and their purposes.

In this chapter, we introduce different kinds of consent with the help of legal theory. We then propose some concrete suggestions for how consent can be broken down and understood for settings of human-robot interaction, as well as how it can be used to drive both empirical and technical research forward. More specifically, in Section 8.2 we take a deep dive into five different forms of consent defined in common law, exploring, with examples, how they influence social behavior and normative expectation. In Section 8.3 we resurface and briefly discuss how consent issues can arise in robotic application scenarios that are relatively mundane, and

not "normatively-weighty." In Section 8.4, we discuss the implications for the HRI community and put forth four research directions – involving research questions and architectural considerations – that take into account this nuanced understanding of consent. In Section 8.5, we provide some practical recommendations for how HRI researchers can make their results more consent-aware in the near-term.

8.2 Legal Landscape of Consent under the Law of Intentional Torts

8.2.1 Conventions, Norms and Laws

The normative landscape of a society can be classified into conventions, norms, and laws. These govern individuals' behaviors, based on the degree of their normative strength, and the level or type of sanctioning for their violations. For example, a behavior or conduct that is a convention suggests a stable pattern of behavior [HLM+17], a societal regularity like people choosing to walk on the right, even when indoors. There may or may not be consequences for violating a convention. Even if they are violated, the wrongdoer is typically presented with a mild push to conform or a small dose of scoffing. Certain types of wrongful conduct can rise to the level of a norm violation. We can think of norms as conventions with an added level of sanctioning, in a stricter sense than the scoffing associated with conventions. Sanctions for norm violations can be penalizing and are intended to change the behavior of the wrongdoer. For example, walking into a Hindu temple with shoes on is a serious norm violation that can incur social sanctioning and potential banishment from the corresponding social group. Behaviors achieve their highest normative level when they are codified into law. Laws deter wrongful behaviors that society has deemed to be harmful to the basic functioning of the society. Without laws, society cannot run. Because laws are codified and interpreted repeatedly over the course of centuries, the legal landscape has attained a degree of maturity and sophistication that can serve as a guide to situate less official normative elements like conventions

and norms. That is, exploring and dissecting the legal landscape can provide us with some ammunition when we tackle less normatively charged situations, ones that are more common in everyday life. Below, we explore specifically the notion of *consent* from a legal standpoint, but note that the framing applies to conventions and norms as well.

In particular, we will primarily focus on the legal landscape of torts. In addition to tort law, the doctrine of consent heavily features in criminal law and contract law. In all three legal systems, there can be legal action for wrongful conduct, and consent can play a role in modulating the repurcussions of the conduct. Contract law is concerned with cases where individuals can negotiate and craft their own rights and reponsibilities towards each other. Most of our social interactions, however, are largely non-contractual, and tort law holds individuals legally accountable for the consequences of their actions in these cases. Tort law developed through judicial decisions in common-law courts, and later was codified into statutes and laws. Like tort law, criminal law is codified in statutes and laws and is aimed at addressing wrongful conduct. Still, criminal law is different in at least a couple of regards. First, tort law views certain acts as wrong in the private sense, one for which a defendant will owe damages to the victim. Criminal acts, on the other hand, are deemed to have wronged not just the victim, but the society as a whole, and the wrongdoer is deemed to be deserving of punishment. Second, while crimes can also be torts, criminal liability requires the defendent to have a more culpable state of mind than does tort liability.[1] Here, we are interested in exploring how consent can (or cannot) mitigate behavior that is deemed wrongful in the broadest sense – everything from the slightest deviations from customs to the most egregious violations of our moral code. To do so, we think that the law of intentional torts allows us to best begin this exploration, as it covers wrongful conduct broadly without also having to tackle issues of contractual meeting-of-the-minds, mens rea, and punishment.

As noted earlier, torts are a legal wrong committed upon person or property

[1]There are other differences between criminal law and tort law including who can bring a case - individuals or government, where can a case be brought - civil or criminal, and what are the possibile penalities.

and, unlike contracts, require no prior agreement between parties. Every member of our society is obligated to pay damages if their actions harm others. One variety of torts, known as intentional torts, is particularly relevant and covers those situations in which intentional actions cause harm. There are several types of intentional torts, including harmful contact (battery), apprehension of harmful contact (assault), unlawful possession, use and alteration of personal property (trespass to chattels, conversion), restricting personal movement of others (false imprisonment), unlawful entry into the land of another (trespass to land), and intentional infliction of emotional distress. A crucial doctrine applicable in all types of intentional torts is the notion of consent, which can alter the morality of another's conduct [Hur96] and potentially permit otherwise wrongful conduct. The doctrine is deeply rooted in history and stems from the Roman maxim *volenti non fit injuria*, latin for "to a willing person no injury is done."

Consent can convert what would be "trespass into a dinner party, a battery into a boxing match, a theft into a gift, and rape into consensual sex"[Hur96]. Consent can not only convert a wrongful behavior into right, but also can grant another the right to do wrong. In this latter sense, it is less about flipping the normative valence of an act to make it right, and more about granting permission to perform what is ordinarily a morally wrong act. In both cases, consent has a tremendous amount of normative power and is an important concept that is central to every interaction. There are several forms the legal notion of consent can take and we explore each below.

8.2.2 Actual Consent

Actual consent is the "gold standard of consent" and is a person's subjective willingness to permit tortious (wrongful) conduct [Sim17]. It is generally considered to be the case that consent is limited to the person's mental state and does not also include the communicative act or signal. That a person demonstrates or evidences their consent via language, gesture and other means is separate from their truest "consent," which is their subjective acquiescence.

A holds out her glass and asks B to take it and bring her some water.
B takes the glass from A. (2)

Here, A has consented to what would otherwise be trespass to chattels, or wrongful taking of property. Now, B's actions would need to cause some damages for them to be tortious, but we will set aside other such legal issues (damages, causation, intent etc.). Instead, we will focus on conduct that could be wrongful, and explore how consent impacts it.

We can distinguish two forms of actual consent: *express* and *inferred*. Express consent refers to situations when the consent is demonstrated explicitly either orally or in written form (i.e., linguistically). Inferred consent is when the subjective willingness is inferred from the person's conduct. (1) was an example of inferred consent, while (2) below is express.

A says to B "You may take my plate." B takes A's plate. (3)

The statement "you may take my plate" by A literally grants permission to B to take away the plate. A slight variant of this if A says "take my plate" leaving out the permission granting part. Note that while the statement "take my plate" does not literally state "you are permitted to take my plate", this permission can be directly inferred from the imperative form (as it does not make sense to instruct someone to take the plate without allowing them to do so – this is the crux of the old deontic principle ascribed to Immanuel Kant that "ought implies can").Utterances where the utterance semantics does not directly express the intended semantics are called "indirect speech acts" (ISAs) and are widely used in social contexts for proper conduct, often to meet politeness norms (e.g., [BS13]). We will discuss more about ISA's below.

Returning to express consent, one distinction not often drawn in the legal literature is between express consent that is direct versus that which is indirect.

A says to B, a bellhop, "Can you pick up my bag, please?
It's quite heavy." B picks up A's bag. (4)

Here, A was clearly consenting to B's actions, but it was certainly not direct in the sense of the semantics of the utterance. B needed to determine that A's utterance was in fact a command or at least a request for B to perform a specific action. ISAs of this sort can be highly conventionalized, and thus it is possible B interpreted the specific linguistic form just as he would the direct utterance "pick up my bag." Crucially, because consent is a mental state, we must consider the possibility that direct and indirect expressions are possible.

The degree to which a dialogue can be indirect is exemplified in (4), below.

A arrives at an expensive restaurant and is greeted by B, the Mâitre D'.
B welcomes A and then walks over close to A and says "may I?." (5)
A extends his arms. B takes off A's coat.

Here, A actually consents to B's contact and taking his property. Not only is consent signal indirect, even B's utterance soliciting the consent is indirect. In fact, B's utterance does not even reference the particular act for which B is seeking consent. Yet somehow, A has deciphered what act B wants to perform and consented to this act. The story would have been quite different if B, after asking A's permission then proceeds to do a pat-down. This would be acceptable in airport security, but not in a restaurant.

8.2.3 Apparent Consent

A second, and lesser, form of consent is one in which the person does not actually consent, but the actor reasonably believes that the person actually consents. In these cases, the person has not formed the subjective willingness to permit the conduct. Tort law requires not just that there be no actual consent, but also that there be no *reasonable belief* that the person consents. Unlike actual consent, which transforms or negates the wrongfulness of the act, apparent consent serves to grant permission to an act that still remains morally wrong.

A (never been to a doctor) visits a doctor. B, the nurse, touches A to take her blood pressure. B does not reasonably know that A is averse to physical touching. (6)

A gets up from her table at an expensive restaurant to go to the bathroom. She places her crumpled napkin on the table by her plate. B, a waiter, comes over and folds her napkin. A has previously never been to expensive sit-down restaurants before and find this a bit strange that B touched her napkin. (7)

In (5) (a variant of [18991]) and (6) B cannot rely on actual consent, because A was not actually consenting, but B might be able to use apparent consent (or possibly even presumed consent, discussed below) to preclude liability for battery in (5) and normative sanctioning in (6). A may or may not have liked her napkin to have been folded in (6) but whether she objected to it or even just did not consider it, we can argue that she did not actually consent to it in either case.

The question of whether the actor can "reasonably believe" that the person consents can be based on evidence obtained from the person's conduct, others' conduct, or customary norms at play.

At the end of a successful business meeting, B reaches out and shakes A's hands before A can react. Unbeknown to B, A objects to touching and shaking hands when greeting people. (8)

A and B are moving about in a busy restaurant. B bumps into A as he was walking to the bathroom. A objects to touching. (9)

A is standing in front of B in line in a coffee shop. Believing A to be his wife based on her demeanor, pocket book she was carrying and her outfit, B gives her a surprise hug from behind. A, however, turned out not to be B's wife and objected to touching. (10)

In (5), (7) and (8) the evidence for B's arguably reasonable belief that A was actually consenting may come from what is customary in a doctor's office or in

a business meeting or in a crowded space. However, if in either case B knew that A objects to being touched, then B cannot rely on apparent consent to preclude liability. In (9), the evidence for B's reasonable belief came not from custom but directly from A's conduct and appearance.

8.2.4 Presumed Consent

We can next consider the case when the actor knows that the person did not actually consent, but reasonably believes that the person *would* actually consent, if asked. This is a counterfactual form of consent. This is different from apparent consent, where the actor reasonably believes that the person actually consents. This form of consent is typically, but not always, encountered in emergency situations when it is infeasible to obtain consent.

> *A, a child, is chasing after his ball runs into oncoming traffic on a busy road. B, who's nearby, notices this and pushes A out of harm's way, but in doing so A breaks his arm.* (11)

> *A is standing on the street corner. B walks up to A and taps her on the shoulder and asks her for directions.* (12)

In both (10) and (11) B is not liable to A because of presumed consent. In (10) the situation is emergent whereas in (11) it is not. Presumed consent has a high burden, additionally requiring a justification for the contact. For example, the actor must show that the contact was minor and was in fact customary in the community.

8.2.5 Constructive Consent

Often known as "implied-in-law" this form of consent is rare, but quite extreme. It considers conduct that causes socially justifiable minor acts that not only *not* require consent, but in fact can be performed *despite explicit objections* to the act and expressions of non-consent. Here, the standard is an objective one of whether a reasonable person would consent under the circumstances.

A drives to work and arrives at the scene of an accident. B, a police officer, orders him to wait in the car while the scene is cleared.
A demands that he be allowed to turn around and leave the scene.
B refuses fearing safety risks.
(13)

B is not liable to A for false imprisonment, a variety of intentional tort, because of constructive or implied-in-law consent. This form of consent is controversial particularly because of its blatant disregard for the person's actual objections.

The concept of consent permeates our legal system beyond intentional torts. Consent is a critical defense to certain criminal conduct and plays a role in mutually beneficial contractual agreements. One interesting twist in criminal law is the existence of a category where consent is not a defense. That is, even if the person granted the actor consent, the conduct might still be deemed to be wrongful and the actor culpable. For example [19880, 20112]:

A and B engage in a sex act in which A consents to B using a dangerous weapon during sex. B uses the weapon on A and A is harmed.
(14)

Here, B might be criminally liable in the Commonwealth of Massachusetts as consent is no defense to a charge of battery and assault with a dangerous weapon.

8.2.6 Reluctant consent

The framing of consent as a defense (to a tort claim or a criminal charge) does not consider how fervently one gives consent. If there is legal basis in evidence for one or more forms of the above consent it can be sufficient. However, exploring the amount of enthusiasm shown in regards to consent could have normative implications even if it does not have legal ones.

A enters a taxi driven by B. B takes an unusual route that involves driving through a part of town that A associates with past emotional trauma.
A says nothing.
(15)

By not objecting to this route, B might be able to rely on a subsequent

defense of apparent consent when faced with an intentional infliction of emotional distress claim. It might even be the way most taxi drivers drive to the location that A intends to reach. However, A's consent is absent and, at best, reluctant. The level of A's consent might be evidenced by facial expressions suggesting it or by other gestures like looking away from the windows or closing eyes. Normatively, it seems B should take this reluctance into account and explore an alternative route. The norms at play include business aspects of ensuring that the passenger is kept happy, the human decency aspect of not having an interactant relive a past trauma and so on. Thus, the legal framework, while giving us a grounding for extreme behavioral violations, might be insufficient. We might need to consider lesser norm and convention violations that are especially evident in cases of reluctant consent.

A thorny issue in (14) is whether B had the duty to present other alternatives. The legal system is not entirely clear on this issue and some scholars have compared reluctant consent with fullness (or partial-ness) of consent and their connections to the doctrine of the assumption of risk. What risk does A assume in taking B's route? If B presented A with a second less-traumatic route, and A still reluctantly chose the first route, then the doctrine of assumption of risk may play a role in clearing B. One takeaway here is that the consenting process is much more complicated than a single interaction and typically involves a complex mix of duties and responsibilities of the actor and the mental state and behaviors of the person who arguably is deemed to have consented.

8.2.7 Other Issues

We have only scratched the surface of the intentional tort landscape around the notion of consent. There are numerous other issues around whether a person is competent or has the capacity to even grant consent. The law has established that for consent to be valid the person must have capacity to consent. Capacity is typically gauged by age and mental competence. Children, those suffering from mental disorders, and intoxicated persons are deemed to not have the capacity to consent. This means tortious behavior will not be defensible even if consent is obtained from

those without the capacity. Consent that is granted can also be revoked at any time. Consent can be conditional in that it is ineffective if there has been a failure to meet conditions that the consenter has imposed. Consent must also be voluntarily given and not coerced or obtained under duress.

Beyond intentional torts, notions of consent appear as another related legal principle in torts of negligence known as "assumption of risk." Assumption of risk is often treated as consent to conduct that is merely negligent [Sim06, Sim87]. The doctrine of assumed risk connects consent with whether there might have been some uncertainty associated with the consequence of the conduct. In the medical context, a related notion known as the doctrine of informed consent is used to receive a patient's acknowledgement that there are risks associated with a particular procedure even if the doctor performed it correctly.

Consent is also a crucial doctrine in criminal law (as noted earlier) and in contract law. Unlike tort law, consent in contract law is centered not around internal subjective mental state, but external performance. Moreover, in contract law, consent serves as a proxy for evaluating whether the parties voluntarily entered into a contract.

Finally, it is worth clarifying that the legal principles in the chapter merely serve as a guide to unpacking an otherwise complex normative landscape in many social settings. We do not require or even suggest that robots be granted any form of legal status as a person or agent. On the other hand, we want to stress that it is insufficient to treat robots as any other form of technology where consent is equivalent to user preferences. Past research has shown that robots create expectations in people that they will be less like passive technologies (e.g., washing machines and cell phones) and more like persons. While we do not take a position on the question of whether robots have "agency" in a philosophical sense, we do encourage thinking about humans in social spaces not in terms of "users" but in terms of "interactants" [Cal10]. Robots, as part of their functions, might interact with other non-users as well. Moreover, it is not enough to express the limitations of the robot to users as users impute agency and capabilities beyond that which the robot has regardless of

prior disclaimers.

Entangled with these foundational legal notions, the nuances of consent can be affected by other aspects of a social context like, for example, gender and culture. Gender and cultural variables associated with interactants and social groups can affect behavior and expectations associated with consent. For example, what is considered presumed consent in one culture might not be in another. Not only will robots need to account for the cultural considerations of the society in which they are deployed, but also potentially the cultural diversity of their interactants within that society. Many communal settings like restaurants and parks might require robots to carefully acknowledge cultural variations among their human interactants.

8.3 Consent Issues in Robot Applications

Consent issues can arise any time there is interaction between agents. As we have seen, these consent issues come in a variety of forms and are highly dependent on the social space in which the interaction is taking place. Thus, although the notion of consent might seem somewhat dyadic, what influences the dynamics of granting and revoking of consent is not merely dyadic, but in fact influenced by the normative expectations of the entire social space. This means consent issues can arise even in scenarios that on the face are not socially or morally charged, as the social setting affects the normative expectations of the interactants. In this section, we will consider two robot applications, in increasing order of "normative-charged-ness" of scenarios in which they are typically deployed: (1) robot vacuum cleaners and (2) robot waiters.

8.3.1 Robot Vacuum Cleaners

Some have estimated that robots account for over 20% of the world's vacuum cleaners. Even if this is a bit overstated, robot vacuum cleaners are ubiquitous. These robots have a range of actuation and sensing capabilities, albeit much more limited than a humanoid robot. They have cameras, IR and laser sensors and gyros

to be able to map out a house and remember what parts of the house have been cleaned and what remains. They also have other features to avoid bumps, recognize no-go areas in the house and schedule cleaning sessions. Many are beginning to come equipped with WiFi capabilities and connections to Alexa and Google Home voice-activated personal assistants. These robots are not simply passive devices that provide a service, they are mobile robots that, numerous studies have shown, cause humans to impute agency and capabilities to, beyond even that which the robot actually possesses [SGGC07, dG16, Sch11].

The robot vacuum cleaner's primary designed function is to vacuum and clean the house, and not necessarily to interact socially with humans. Even so, before, during and after the robot performs its cleaning duties, it can and will interact with humans because it is deployed in an otherwise social space. It's owner might interact with it to decide appropriate cleaning schedules so as to take into account the presence or absence of other members of the household. The robot vacuum cleaner might inadvertently interact with non-owner humans who might be present in the space the robot is cleaning. Humans might direct the robot to go elsewhere or take an alternative cleaning route to avoid being disturbed. During its operation, the robot is also mapping the house and possibly recording video as it learns about the layout of the house and habits of the humans living there. Consent issues can arise quite easily in many of these interactions - robot getting in the way of humans moving about the house, cleaning parts of the house that are not necessarily no-go areas, but temporarily off-limits, cleaning near children's toys, some of which might be small parts, listening to private conversations in the name of listening for human commands, and recording video in private areas of the house. In many of these cases, the humans may not have actually consented, but can the robot assume there is apparent or presumed consent?

8.3.2 Robot Waiters

We have seen in the examples presented in Section 8.2 that consent issues can arise quite frequently in restaurant scenarios. Robot waiters constantly interact with their

assigned customers, other customers walking in the restaurant, chefs, and managers. During these interactions, the robots through their appearance and functional (actuation and sensing) capabilities generate a set of expectations in the interactant humans that together establishes the social space for that particular interaction [RR11].

Distinct from robot vacuum cleaners, however, robot waiters are designed to provide a service that is inherently social, but not necessarily high-stakes or weighty. They direct customers to their seats, take orders, deliver food, check for comfort, and are meant to generally address grievances and comments issued by the customers. Each of these interactions involves entering personal spaces, touching humans, giving and taking possession of objects (some belonging to the customer). Most, if not all, of these interactions are consent-driven, in that some form of implicit or explicit consent was given or presumed, otherwise these behaviors would be problematic.

For example, consider the situation where the robot waiter is tasked with removing plates from a customer's table once they have finished eating. Customers might explicitly call out to a robot and ask for their plate to be removed. They might actually let the robot waiter know that they have finished their meal, as an indirect way to signal their desire for the plate to be removed and a check to be delivered. More often than not, customers will not make this request explicitly or even indirectly. Instead, they will provide implicit cues to suggest that they are done with a particular part of the meal and that the waiter *can* take away their plate. They issue this consent by either arranging their forks and knives a particular way, or push their plate away a bit, or even lean back away from the table as the waiter approaches. It is also possible that the customer is not really thinking about the waiter taking away their plate, but would not mind if the waiter did so. This might be conveyed via a smile that is offered after the waiter has already leaned in and grabbed the plate. The waiter may have assumed the customer was done with their meal because they had not been eating anything from it for a while. Robot waiters need to be able to read these subtle cues. Robot waiters must also be able to read cues in which the customer does not want their plate taken away from them, even if

it seems as if they are done, or they have inadvertently placed their fork and knife in a "finished" position. If the robot waiter does not pick up on refusal hand-gestures or on the customer's grabbing of a fork to show they are still eating, it can negatively impact the service and experience. It is worthwhile to note that regardless of the aesthetic appeal of the robot or of its ability to display pleasant emotions, taking away a customer's plate while they are still eating without their consent (in one of the forms mentioned earlier) can ruin the interaction.

This is just one example of a scenario where a robot's behavior, which would generally have been considered preferable, is suddenly deemed unwanted. Such a robot must be able to recognize the corresponding consent-cue to avoid the entire interaction from being impacted. Other such examples of behaviors that might require a consent-cue include: friendly touches of the customer's shoulder, filling or not filling water or wine, approaching the table to check on the customer's experience, picking up an item the customer may have dropped on the floor, clearing a path between tables by moving customer's garments, bags and chairs or requesting customers to make way, helping a customer take off their coat, and many more.

8.4 Research Directions for HRI

In this section, we explore possible research directions for HRI as the legal structure presented in Section 8.2 allows us to think about consent more deeply. But, how can this legal structure help with designing better human-robot interactions? We specifically consider four research directions or themes that the HRI research community is best positioned to address: (1) when consent is applicable, (2) how consent can be detected, (3) what happens when snap judgments must be made, and (4) what the different roles for a robot are when it is fully integrated into the normative fabric of society. For each theme, we will raise some research questions and provide architectural considerations that the HRI community might consider.

8.4.1 Applicability of Consent

When does a behavior need consent? Presumably, there are behaviors that robots can perform that do not need consent. For example, a vacuuming cleaning robot cleaning a kitchen floor might not need consent for choosing to follow one (random) cleaning path versus another (random) cleaning path. Even in more socially-charged settings, it would be unnecessary for say an elder-care service robot to have to secure approval for every single motor movement. This would be burdensome and possibly even negatively impact the interaction itself.

Current HRI research has provided some guidance into what behaviors are preferred versus those that are unpreferred. The angle of approach example presented in Section 10.1 is one such contribution. We can, at a first approximation, note that unpreferred behaviors will probably need to be consented or at least need some sort of justification (viz. apparent consent, presumed consent). Also, it is possible that in many cases even preferred behaviors could require consenting. Behaviors are situated in a dynamic environment and their normative valence and the applicability of consent can, therefore, shift in time. Consider this example:

> *A and B are sitting at a shared library table that can seat two. While studying, B moves some textbooks around pushing several books over to A's side of the table.* (16)

B's behavior caused an intrusion into A's space. B's behavior, however, did not start out that way. It was when B's books crossed the imaginary halfway line on the table between A and B that B needed consent. The implicit norm at play was the idea of a fair partitioning of a shared work space. The norm became activated once B's book crossed this imaginary center line.

Or consider this service example:

A is a patient at a hospital and is recovering from a surgery in his assigned room. A receives guests and wellwishers who bring him cards, flowers and some of A's favorite take-out Chinese food, which they share during the visit. (17)
B, a janitor, is tasked with cleaning A's room and throws out not only the Chinese food containers, but also some of the cards.

Once allowed into the room, B's behavior of clearing the dirty food containers does not need consenting from A, but clearing the adjacent greeting card might. The underlying norm hinges on the agent being able to detect and consider questions of ownership, possession, and the use of space, in addition to the normative affordances offered by the objects in the environment. The greeting card serves the purpose of lifting A's spirits and therefore is an important component of the environment. B's action of clearing it might eliminate this positive utility. Note, it might be entirely inappropriate to even ask for consent to clear the greeting card as that would suggest B's normative proclivities.

8.4.1.1 Research Questions

When does a behavior turn from being consent-free to one that requires it? Are there some behaviors that require robot consent, but not human consent? Are social interaction cues of proximity, eye gaze, and so on also cues suggesting that consent may be needed? Who should be consenting in a particular situation and if that person is unavailable, should the behavior be postponed? Do humans expect robots to have the same level of understanding of consent as they do with other humans? Do they hold robots to higher standard, i.e., more conservative, when it comes to asking for consent? A popular maxim used in connection with consent in sexual acts is – "yes means yes, no means no." – do we expect the same understanding from our robots?

8.4.1.2 Architectural Considerations

One important shift in architecture design that is prompted by consent processing is that even what are otherwise considered to be automatic behaviors in robots (e.g., approach and obstacle avoidance behavior, eye gaze and blinking behavior, posture and gesturing, etc.) are subject to modulations based on consent. This, in turn, requires the architecture to have built-in suppression mechanisms that can be triggered by changing consent status and interrupt, suppress, or alter these reflex-like behaviors. In addition, the architecture requires a knowledge base with common sense knowledge about the different consenting conditions and the appropriate modulations. For example, it will require general knowledge about possession of property, what actions are and are not allowed with another person's property, and under what circumstances the robot can dispense with certain rules and norms because there are role-based expectations that trump them. In the above example, even though the food container was technically owned by patient A, a janitor robot B would have had to make the inference that based on its role to clean up used items – and an empty food container was a used item – the ownership of the container was no longer an issue and it was, in fact, expected from B to take the container away from A. This is a case where two norms clash, the norm of following ownership rules versus the norm of fulfilling one's duties based on one's roles. The challenge then is to detect that the role-based norm of cleaning up the container trumps the ownership-based norm of not touching it precisely because it implies consent by the owner to do so. Recently, research progress in explicit norm representations [SSA+17, SSM17], norms relating to object affordances [SS16d], norms relating to ownership [XT19] and other prerequisites for consent reference what social norms might be present and how they legitimize behavior, which in turn might suggest how they are triggered and modulated by consent.

Detecting shifting consent-dependencies of behaviors is particularly challenging from a technical standpoint. Detecting consent applicability shares many similar technical aspects to detecting consent itself. The robot will need to detect a variety

of cues from its interactant, others in the environment, and other animate and inanimate objects that constitute the social space. A technical designer must consider the level of symbolic abstraction for a representing a behavior before it qualifies as one applicable for consent. Clearly, step-wise motor movements and joint angles might be too specific and not necessarily applicable. On the other hand, higher-level behaviors of swinging arms and even higher-level conduct of picking up and manipulating objects might be candidates for consent-dependency. It is also possible that the robot might need to consider the nature of objects it is handling (e.g., knives versus a bouquet of roses) when performing certain behaviors.

8.4.2 Detecting Consent

At the core of consent-related considerations that we have discussed in this chapter is the question of how and what consent cues are detected. At first glance it seems like language, gestures, facial expressions, eye gaze, approach, proximity can all serve as valid vehicles to communicate one's subjective intent of actual consent or objection. But, how can a robot detect consent expressed through an ISA, or through the usage of objects in the environment? Consider this example:

> *A was sitting at the bar getting a drink that he already paid for. A drinks most, but not all, of the drink and then leaves the bar. B a waiter did not see A leave. B takes away A's almost empty mug.* (18)

Presumably, B's conduct was normatively appropriate given that A might not have objected to the mug being taken away. Note, the mug actually belongs to the restaurant (leaving aside the issue of who the remaining drink belongs to), so there is no legal consequence here. We nevertheless use the legal framework to inform our discussion about consent norms and consent conventions. So, even if A had just stepped out for a smoke or to take a call, B's actions might have been justified. Now, contrast this example with one where A leaves his bag on the seat and goes to the restroom. Here, B's actions of taking away the mug might not be justified. A might object that he had not finished his drink and his intentions were

made clear by him leaving his bag on the chair.

There are many different types of consent cues. Some cues are *consent-granting*, in which case they have an illocutionary force of conveying permission to a behavior. Alternatively, some cues are *consent-denying*, in which case they have the force of conveying explicit non-consent to a behavior. In some instances, the person is not yet ready to signal, but they will eventually. These *consent-withholding* cues suggest that the actor may return at a later specified or unspecified point to inquire about the consent. For example,

> *B, a waiter, is about refill A's wine glass. A gestures with a downward facing open palm and says, "not yet." B retracts.* (19)

Here, A did not say to never fill his glass, instead withheld consent until some later point. Now, there is a subtlety here where A might have instead just said "no", which might have been a consent-denying cue, or if prior consent had been granted, a *consent-revoking* cue, taking away a prior consent. Finally, the consent itself might be conditioned and expressed via a *consent-conditional* cue of A saying "only if it is a 1991 Argentinian Malbec" suggesting that refilling is only acceptable for a particular variety of wine.

These different types of cues not only convey whether or not consent is given, but also specify which behavior they refer to. In (18), when A says "not yet" they are not withholding eventual consent for B pouring the wine on them. The content of the consent is difficult question that poses many challenges.

The legal landscape also suggests that for succeeding on a defense of consent, the actor must also prove that the person has the capacity to consent, is consenting voluntarily and has all the information necessary to decide. The latter of these elements is the basis for the doctrine of informed consent, common in medical settings.

8.4.2.1 Research Questions

How to detect social and environmental consent cues? Are there certain cues that transcend situations and universally signify consent or non-consent for any behavior?

Is it enough for the robot to detect and then immediately act on a consent cue or should it ask for confirmation? What is the role of uncertainty in detection? Crucially, what if the cue was not only unclear in a noisy environment, but also ambiguous in its scope and content? If the robot is unable to glean actual consent, should it explicitly ask the human or should it be satisfied with inferring apparent consent?

8.4.2.2 Architectural Considerations

To be able to detect cues, the robotic architecture must be able to first process and integrate percepts that can serve as cues obtained through various sensory modalities. Consider linguistic cues, which, although they seem like the clearest type of cue, can still be a technical challenge to process. Here, processing consent involves a degree of intent recognition and processing of potentially indirect cues (via indirect speech acts). Architecturally, it might not be enough to implement statistical natural language processing and speech recognition systems that do not have a way to represent contextual variations in the illocutionary and perlocutionary force of speech acts. A way around this issue would be to explicitly design an architecture that allows for performing cue-based inference. An approach might be to use a rule-based strategy where implication rules can represent consent-cues as antecedents and behaviors as consequents. Cues then trigger rules that determine appropriate behavior. The normative propriety of the behavior can then computed via logical modus ponens based on the perception of the relevant consent cues. A variant of this rule-based approach can be seen in indirect speech research ([BS13]). This sort of architecture and knowledge representation has also been shown to be amenable to non-statistical, instruction based teaching and learning via natural language.

In addition to linguistic cues, the robot may have to perform inference from visual percepts as in (17). In such cases, the robotic architecture must be equipped to not only recognize objects, but also generate more holistic representations of a scene. This means object properties need to be inferred and, in dynamic environments, objects and their properties will need to be tracked. These object representations

will then need to be integrated with other modalities including the linguistic cues noted above in order to generate a coherent *story* for whether or not the robot has deemed there to be consent in a particular situation.

But that only covers cases of actual consent. The cues in actual consent are typically limited to those that result from the conduct of the person consenting. However, in apparent consent situations, the cues can come from not only from the person who is the target of the behavior in question, but also from others and from customary aspects of the situation.

The robot must also be able to understand the interactants' capacity to consent and ensure that they are provided with all the information they need to make their consent decision. Robots working with certain vulnerable populations – children, elderly, substance-abusers, mentally disordered – must account for these aspects.

8.4.3 When Consent Does Not Matter

What the legal landscape has shown us is that there are situations where an otherwise unconsented behavior is appropriate (presumed consent and constructive consent) and consented behavior is still inappropriate (consent not a defense to certain criminal charges).

In situations like emergencies where consent is presumed, the law places a high burden on those looking to use this principle in an affirmative defense to an intentional tort. Presumed consent falls squarely below actual and apparent consent and has limited scope. But, these are situations in which our robots might find themselves, possibly even more frequently than imagined. Search-and-rescue robots, robot astronauts, medical robots and robots in many emergent situations must take quick action in order to prevent a larger harm.

In certain cases, sex robots engaged in BDSM activities might have received consent for certain acts, but the law has held that performing these acts (particularly those with dangerous weapons) can still remain a criminal charge. On a less egregious note, sex robots that bring spontaneity and creativity into a sexual encounter (e.g.,

a surprise tickle) might need to rely on presumed consent to provide a positive social interaction.

8.4.3.1 Research Questions

Should a robot ever presume consent? If a robot does presume consent, should its scope be kept limited to a few specifically designed behaviors. Presumed consent is particularly controversial and it, combined with constructive consent, starts tearing away at personal autonomy. If a person's consenting of a behavior is taken for granted or presumed given or worse, ignored, is there not a taking away of a moral right to autonomy? Should we give robots this privilege? What if it ends up helping us? Moreover, do we humans expect this from our rescue robots? Some in the HRI community have explored the role of supererogatory actions (as in (10)) and whether we expect a street cleaning robot to go out of its way to save a child from an oncoming car. The technical challenges for designing such a robot share many aspects of cue detection and norm representations noted earlier. In cases where consent is not a defense, should robots be designed to never perform these behaviors? How can we balance robot creativity, spontaneity and risk-taking behaviors (all of which might have positive social value) with normative propriety? Can the robot presume consent if it enhances a social experience?

8.4.3.2 Architectural Considerations

While the question of whether or not consent matters in certain special situations is largely an empirical and philosophical one, we can still evaluate what architectural capacities a robot must have. Uniquely, a robot must be equipped with suitable architectural components to ensure that it does not become a passerby, or bystander, during critical situations. Thus, it is important for there to be coordination within the architecture for integrating cues suggesting an emergent situation and the ability to reason against norm expectations. Crucially, such reasoning is required as the robot needs to deduce that a norm violation might be necessary in order to achieve a greater good. The architecture will need mechanisms to reason with multiple such

mutually inconsistent and conflicting norms, all in real-time under extreme time-pressures imposed by the emergent situation.

8.4.4 Robot Roles

Thus far, we have looked at several examples in which we examined the propriety of actions of B in interactions with A. While it is natural to think about the role of a robot as B, and the human as A, in reality, the role of the robot could be quite varied. Here, we can explore some of the alternative roles the robot plays in an interaction.

8.4.4.1 Research Questions

First, we can consider the robot as a consenter (or protester), i.e., A. The notion of robots in this role raises many questions about robot rights and autonomy [Dar16]. Should the robot be able to consent at all? A narrow view of robot rights would suggest that robot consent does not matter. This view, in turn, might increase the scope of presumed consent or constructive consent in robot interactions. Can we then abuse a robot?

B verbally abuses A a personal assistant robot by shouting expletives. (20)

Some might argue that A's consent towards this behavior is irrelevant as A is a machine. However, the counter argument to this view is that B's verbal abuse sets the wrong normative tone in the environment and therefore should be seen as wrongful. There may not be emotional distress for A, but there could be for another, say C, who is watching this interaction.

We can bring in another interactant C who might be a third-party to the interaction. C could play the role of sanctioner and function to right normative wrongs in an environment. For example, C could be a manager of a restaurant in which B a waiter is serving A a customer. Alternatively, we can think of C as a sanctioner role that could be played by A themselves.

In such a role, we must ask if it is okay for the robot to call-out and sanction normatively harmful behavior (see [Fes18] for Alexa pushing back against abuse). In (19) could the robot defend itself and protest A's abuse? Should the robot do such a thing?

Another role a robot could take is not that of a moral arbiter, but that of the elusive "reasonable person". In much of the legal scholarship one aspect that has been at the center of many debates is the objective standard of a "reasonable person", one who exercises average skill and judgment and can objectively assess a situation and arrive at the correct solution. Many legal scholars have argued that such a person does not exist and is, in fact, by no means average [Gar15]. Nevertheless, robots might present themselves in the unique position to serve as such a reasonable person. Can these reasonable robots help resolve conflicts by performing an unbiased and objective analysis of the facts and situation?

8.4.4.2 Architectural Considerations

The architectural capabilities we have discussed thus far apply to the robots that serve as sanctioners or reasonable persons. Of particular note, these robots must be able to detect norm violations, and so also have architectural components to compare normative cues with norm-based expectations to detect norm violations. Robots in these varying roles also have the function of deciding if the justifications provided by parties are acceptable and as such must have mechanisms to infer consent by exploring the purposes of actions. Investigations into human-robot teamwork have even included their possible role as repairers of social conflict [JMH15]. Robots must also maintain honesty [HFA+15] and sometimes even keep secrets [KJKI+15] in order to fulfill the responsibilities [Asa07] imposed by their roles. This also means that robotic architectures serving any role must be able to represent various supporting aspects of the situations – ownership, possession, resulting permissions and obligations and whether or not agents have the capacity to consent and have done so voluntarily with the needed information. The architectures must also be able to determine precedence relationships to help enable consent inference.

8.5 Near-Term Next Steps for HRI

In the previous section, we discussed four research directions (or themes) that are opened for the HRI community by the more nuanced consideration of consent. Critically, we proposed research questions and technical (architectural) design issues that will need to be addressed if we are to design fully consent-aware robots. In some sense, we presented a research vision for normative HRI that is focused on much-needed technical advancements that are likely to be longer-term. But, what can HRI researchers start doing today? We do not yet have all the technical capabilities to build the architectural components needed for fully consentual agents. However, as we have discussed in Section 8.3, robots are already deployed in many real world social settings and will need to be able to handle at least some partial set of consent cues. How can HRI researchers design and evaluate more consent-aware interactions with these current robots? In this section, we provide some recommendations for how current HRI researchers can adapt their design and evaluation methodologies to begin accounting for consent. We expect that by doing so, this will guide future technical developments in the right direction.

8.5.1 Early-Stage Interaction Design

HRI design is a truly collaborative endeavor and tackling some of the consent issues might need to go beyond architectural modifications of robots. That is, it is insufficient to only take a robot/technical-centric view and place all the burden of consent-management to the robot and its underlying A.I. Accounting for consent must begin at the earliest stages of interaction design. Interaction design involves considering not just behaviors in isolation, but the generation of sequences of behaviors and how such sequences influence and are influenced by particular application scenarios. Behaviors could include both dialogue and non-dialogue (body movements), and interaction forms the overall specification of the set of behaviors that technical researchers must build into their robots. Interaction designers can proactively consider consent when designing specific interactions by asking the research

questions we presented above *early* in the design stages of storyboarding and prototyping. Asking when and how consent will apply to particular scenarios by designers at this stage will likely prove crucial when designing the specifics of the interaction itself and help define the set of robot capabilities. As behaviors are evaluated against a scenario and flows constructed, interaction designers can ask critically *when* and *why* certain behaviors are acceptable or not, clueing them into whether or not consent applies in a particular situation. For example, asking when and why it is okay for a robot waiter to fill an empty water glass raises many interesting questions about the situation and the nature of the social space that might need to be answered before a suitable interaction can be designed. The answers might suggest entirely different interactions for an expensive restaurant versus a fast food restaurant. A deeper inquiry at that point might shed some light on what factors might turn acceptable behaviors into unacceptable ones and vice versa. For example, customers in expensive restaurants might provide consent cues that are much more indirect and implicit than those at fast food restaurants. Similarly, the interaction designer could attempt to model the mental state of users in the scenario to consider what these users would consider reasonable deviations from acceptable behavior. If the scenario involves emergencies, interaction designers will be able to directly incorporate notions of presumed consent. For example, when a robot detects that a human is about to get run over, it can assume presumed consent for pushing the person out of harms way, even though pushing normally would violate a tort.

8.5.2 Experimental Evaluation and User Studies

When robotic systems are introduced into a real world application scenario, they can have a significant impact on usability, user experience, and social aspects like impact and acceptance. Prior to such an introduction, a strong evaluation methodology is needed to enable fair comparisons of competing HRI systems along various dimensions including efficiency, feasibility, safety, and social and psychological aspects. There have been many approaches proposed in the literature for the most suitable set of metrics for evaluating robots. A prominent one is by Steinfeld et al.

[SFK+06] who propose a set of task-specific as well as common metrics for assessing a robot's performance. In this comprehensive work, they discuss some socially-driven metrics (persuasiveness, trust, engagement and compliance), but these are largely discussed in connection with robots whose primary function is to interact socially with people. As we have discussed, socially-relevant consent issues can arise even when the robot's primary function is not social interaction; consent issues can be evoked from the robot simply operating in a social space. The common metrics proposed by Steinfeld et al. do not consider such normative and socio-cultural aspects of a scenario in which the robot is deployed. Thus, urgent future work is needed to develop suitable metrics and measurement instruments that explicitly probe for consent.

In addition, we propose that interaction researchers studying behaviors, but not necessarily studying consent, can and should provide some more guidance about the ecological validity of their findings. They can do this by asking when preferred behaviors become dispreferred and vice-versa. User-study questionnaires already ask questions about user preferences and likeability of behaviors. These can be extended to include questions about dislikeability and what makes a certain behavior inappropriate, and when dispreferred behavior could become preferred. By exploring ways to extend these questionnaires we can study how consent influences existing interaction research findings and provide some new ideas for future research inquiries.

More generally, when robot behaviors are being evaluated for a particular use-case, HRI researchers must consider the nature of normative expectations held by humans present in that use-case, and ensure that their experimental subject-population is representative of those in the real world. Most HRI user studies obtain demographic information either to directly study gender or some other effect relating to the identity of participants, or indirectly to correct for individual differences across a population. We are suggesting here that consent-preferences are an equally important (an possibly biasing or confounding) factor to either study or at least correct for in user studies.

Finally, when drawing conclusions from user-studies HRI researchers must

make explicit the types of consent-related aspects they have accounted for in their experiment. That is, their conclusions and claims about behavioral preferences must be accompanied by the scope and nature of the scenario(s) in which those behaviors are situated. During the peer-review process, we urge reviewers to extend this inquiry and consider how consent or lack thereof can impact these results and push the researchers to make explicit these limits.

8.5.3 Revisiting Past HRI Experiments

The future outlook for HRI research incorporating this rich notion of consent is quite exciting, and we hope the suggestions provided in this chapter will facilitate a smoother transfer of future interaction research results into real world robotic application domains. That said, it is still unclear what we are to do with past and current research results like, for example, research on "robot approach" that we alluded to in Section 8.1. A formidable challenge lies in understanding the scope of the results in not only approach research, but also dialogue, eye gaze, proxemics, touch and other interactionally-relevant behaviors. We know that behavioral preferences can be highly contextualized and dictated by the task at hand. But, it is unclear how the results for particular studies can be generalized beyond the specific experimental setting in which they were designed.

For example, in eye gaze research, HRI researchers have suggested that high levels of mutual gaze express feelings of trust and extroversion, and gaze aversions express feelings of distrust and introversion [AS17a]. However, this is not universally true, as an unwelcome extended eye gaze might be suspect and can adversely impact trust. In conversational contexts, eye contact is seen as more acceptable in emotionally neutral topics but not so when the conversation is embarrassing [AS17a], suggesting that the topic of conversation could be a modulating factor when evaluating robot perceptions. Although there could be other modulating factors in particular scenarios, we propose that consent is a key modulating factor in all HRI settings, and whether or not certain eye gaze strategies are deemed socially acceptable (or any other socially relevant evaluation metric) is dependent on whether the interactants

mutually consented to each others behavioral choices. In situations where there is no explicit consent, the question then becomes if the social setting provides a reasonable basis for establishing some implicit form of consent. Without a reasonable basis for it, a social space that places a higher bar on consent will not allow for prolonged eye contact even if conversations are emotionally neutral. In order to extend the applicability of many such HRI studies into *real* interactional contexts, what we need is a deeper understanding of the social spaces anticipated by these studies. We need a better sense for what sorts of factors need to be present or absent to reasonably establish that either consent can be apparent or presumed in the social space, or actual consent was indeed given.

As another example, consider research in politeness in natural language. HRI researchers first looked at speech modifications as a way to modulate politeness and thereby improve human perceptions of robot helpers. Torry et al. [TFK13] argued that robot helpers create a positive impression when using hedges ("I guess" or "kind of" or "probably") and discourse markers ("I mean" or "like you know"). Their study involved humans observing robots employing those dialogue strategies, not humans directly interacting with such robots. Strait et al. [SCS14] subsequently noted that such speech modification strategies alone are insufficient and other aspects of the interactional scenario like participants' personality, dialogue efficiency and task success also impacted the perceptions of the robot. They explored the relative effects of communication strategy as modulated by interaction modality and presence as well as robot appearance. Their results showed that some prior claims made by Torry et al. do not generalize to real interactional scenarios where other modulatory factors might be at play. Their conclusions leave open the possibilities for further situational factors that might modulate interactions. We believe this is a step in the right direction, and we urge HRI researcher to make more explicit the spectrum of modulatory factors that might be present in a particular HRI scenario. Specifically though, we reiterate that consent is one such modulatory factor, and an extremely important one at that. We believe the effect of consent is critical to evaluating robot perceptions, and on par with other influential factors such as gender effects.

Regardless of where one stands on the relative importance of the role of consent over other situational factors, it is fair to say that since consent can significantly impact all HRI interactions, it must be explicitly accounted for in our HRI experiments. One direction might be to redo some of the more significant past experiments to include some explicit consent measures and possibly even using consent/no-consent as an independent variable in the design. So, the robot-approach results could be redesigned by taking the presupposed consent structure of the context in which they are performed into account. On the one hand this might sound like bad news, but we think it is not so and will actually substantially strengthen those results that survive this repetition, and provide new opportunities for those results that show some variation when corrected for consent. Importantly, such a revisiting of past experiments might be necessary in order to be able to understand what those interaction design results actually mean when those interaction designs are translated into real world applications deployed in social spaces requiring consent-awareness. We might need to go back to the drawing board on some experiments to get additional results for cases not previously considered or conflated due to the lack of a consent dimension in the experimental design.

8.6 Conclusion

In this chapter, we have presented consent as a crucial area for HRI research. Consent often defines human-robot interactions for their entire duration, rather than being just a precondition that is established or refused beforehand. We have outlined levels of consent to show how subtle social cues can change the nature of what social robots roles are and what actions they can do. From several exemplary scenarios, it is clear that the HRI community can extend insights of existing research into richer areas of human-robot interface. Our brief architectural considerations suggest how robotic systems, designed with architectures incorporating rules-based and context-aware elements, could best recognize how consent is sought, maintained, and respected.

As robots enter into the rough terrain of social spaces, consent will unavoid-

ably take on cultural and political valences. Even seen as tools, robots will assume representations of broader values, for better or worse. Societal critiques have stressed the harm of gendered norms of silence and disregard for consent, and steering clear of consent still implicates robotic systems in how they are embodied. This reflects why for broader discussions of AI ethics, including law and policy, HRI research can be a strong voice for otherwise neglected risks and opportunities posed by embodied automation. It can be at the forefront of guiding how human society can more carefully consent to the technologies meant to serve it.

Part III

Sense-Breaking

Chapter 9

Human Problem Solving

In Part I, we focused on sense-making and the role of logical reasoning over different types of knowledge (esp. social norms) as critical to understanding language and perceiving affordances. In Part II, we took a deep dive into social normative knowledge and explored its cognitive, computational and societal aspects. Both Parts I and II established the importance of knowledge-based assumptions to guide intelligence. These bodies of knowledge and the accompanying reasoning mechanisms help us handle ambiguity and uncertainty effectively and will help future AI systems in making sense of the world. However, our knowledge and assumptions about the world can limit us. In Part III, we explore the realm of creativity, when assumptions must be lifted, and sense-making turns into sense-breaking.

This chapter explores human real world creative problem solving and synthesizes a disparate body of literature in Cognitive Psychology and Cognitive Neuroscience. The chapter proposes a model of creative problem solving, which is then computationalized in subsequent chapters.

9.1 Introduction

In the Apollo 13 space mission, astronauts together with ground control had to overcome several challenges to bring the team safely back to Earth [LK06]. One of these challenges was controlling carbon dioxide levels onboard the space craft: "For

two days straight [they] had worked on how to jury-rig the Odyssey's canisters to the Aquarius's life support system. Now, using materials known to be available onboard the spacecraft – a sock, a plastic bag, the cover of a flight manual, lots of duct tape, and so on – the crew assembled a strange contraption and taped it into place. Carbon dioxide levels immediately began to fall into the safe range" [Cas05, Tea70].

The success of Apollo 13's recovery from failure is often cited as a glowing example of human resourcefulness and inventiveness alongside more well-known inventions and innovations over the course of human history. However, this sort of inventive capability is not restricted to a few creative geniuses, but an ability present in all of us, and exemplified in the following mundane example. Consider a situation when your only suit is covered in lint and you do not own a lint remover. You see a roll of duct tape, and being resourceful you reason that it might be a good substitute. You then solve the problem of lint removal by peeling a full turn's worth of tape and re-attaching it backwards onto the roll to expose the sticky side all around the roll. By rolling it over your suit, you can now pick up all the lint.

In both these examples (historic as well as everyday), we see evidence for our innate ability to problem-solve in the real world. Solving real world problems in real time given constraints posed by one's environment is crucial for survival. At the core of this skill is our mental capability to get out of "sticky situations" or impasses, i.e., difficulties that appear unexpectedly as impassable roadblocks to solving the problem at hand. But, what are the cognitive processes that enable a problem solver to overcome such impasses and arrive at a solution, or at least a set of promising next steps?

A central aspect of this type of real world problem solving is the role played by the solver's surrounding environment during the problem-solving process. Is it possible that interaction with one's environment can facilitate creative thinking? The answer to this question seems somewhat obvious when one considers the most famous anecdotal account of creative problem solving, namely that of Archimedes of Syracuse. During a bath, he found a novel way to check if the King's crown contained non-gold impurities. The story has traditionally been associated with the

so-called "Eureka moment", the sudden affective experience when a solution to a particularly thorny problem emerges. In this chapter, I want to temporarily turn our attention away from the specific "aha!" experience itself and take particular note that Archimedes made this discovery, not with his eyes closed at a desk, but in a real-world context of a bath.[1] The bath was not only a passive, relaxing environment for Archimedes, but also a specific source of inspiration. Indeed it was his noticing the displacement of water that gave him a specific methodology for measuring the purity of the crown; by comparing how much water a solid gold bar of the same weight would displace as compared with the crown. This sort of continuous environmental interaction was present when the Apollo 13 engineers discovered their life-saving solution, and when you solved the suit-lint-removal problem with duct tape.

The neural mechanisms underlying problem-solving have been extensively studied in the literature, and there is general agreement about the key functional networks and nodes involved in various stages of problem-solving. In addition, there has been a great deal of work in studying the neural basis for creativity and insight problem solving, which is associated with the sudden emergence of solutions. However, in the context of problem-solving, creativity and insight have been researched as largely an internal process without much interaction with and influence from the external environment [Weg12, Abr13, KB14][2]. Thus, there are open questions of what role the environment plays during real world problem-solving (RWPS) and how the brain enables the assimilation of novel items during these external interactions.

In this chapter, I synthesize the literature on problem-solving, creativity and insight, and particularly focus on how the environment can inform RWPS. I explore three environmentally-informed mechanisms that could play a critical role: (1) partial-cue driven context-shifting, (2) heuristic prototyping and learning novel associations, and (3) learning novel physical inferences. I begin first with some intuitions

[1] My intention is not to ignore the benefits of a concentrated internal thought process which likely occurred as well, but merely to acknowledge the possibility that the environment might have also helped.

[2] The research in insight does extensively use "hints" which are, arguably, a form of external influence. But these hints are highly targeted and might not be available in this explicit form when solving problems in the real world.

about real world problem solving that might help ground this discussion and providing some key distinctions from more traditional problem solving research. Then, I turn to a review of the relevant literature on problem-solving, creativity, and insight, before discussing the three above-mentioned environmentally-driven mechanisms. I conclude with a potential new model and map out its hypothesized neural basis.

9.2 Problem Solving, Creativity and Insight

9.2.1 What is Real World Problem-Solving?

Archimedes was embodied in the real world when he found his solution. In fact, the real world helped him solve the problem. Whether or not these sorts of historic accounts of creative inspiration are accurate[3], they do correlate with some of our own key intuitions about how problem solving occurs "in the wild." Real world problem solving (RWPS) is different from that which occurs in a classroom or in a laboratory during an experiment. It is often dynamic and discontinuous, accompanied by many starts and stops. Solvers are never working on just one problem. Instead, they are simultaneously juggling several problems of varying difficulties and alternating their attention between them. Real world problems are typically ill-defined, and even when they are well-defined, often have open-ended solutions. Coupled with that is the added aspect of uncertainty associated with the solver's problem solving strategies. As introduced earlier, an important dimension of RWPS is the continuous interaction between the solver and their environment. During these interactions, the solver might be inspired or arrive at an "aha!" moment. However, more often than not, the solver experiences dozens of minor discovery events - "hmmm, interesting..." or "wait, what?..." moments. Like discovery events, there's typically never one singular impasse or distraction event. The solver must iterate through the problem solving process experiencing and managing these sorts of intervening events (including impasses and discoveries). In summary, RWPS is quite messy and involves a

[3]The accuracy of these accounts has been placed in doubt. They often are recounted years later, with inaccuracies, and embellished for dramatic effect.

tight interplay between problem solving, creativity and insight. Next, I explore each of these processes in more detail and explicate a possible role of memory, attention, conflict management and perception.

9.2.2 Analytical Problem-Solving

In psychology and neuroscience, *problem-solving* broadly refers to the inferential steps taken by an agent[4] leading from a given state of affairs to a desired goal state [BB09]. The agent does not immediately know how this goal can be reached and must perform some mental operations (i.e., thinking) to determine a solution [Dun45].

The problem solving literature divides problems based on clarity (well-defined vs ill-defined) or on the underlying cognitive processes (analytical, memory retrieval, and insight) [SRE+17]. While memory retrieval is an important process, I consider it as a sub-process to problem solving more generally. I first focus on the analytical problem-solving process, which typically involves problem-representation and encoding, and the process of forming and executing a solution plan [Rob16].

9.2.2.1 Problem Definition and Representation

An important initial phase of problem-solving involves defining the problem and forming a representation in the working memory. During this phase, components of the prefrontal cortex (PFC), default mode network (DMN) and the dorsal anterior cingulate cortex (dACC) have been found to be activated. If the problem is familiar and well-structured, top-down executive control mechanisms are engaged and the left prefrontal cortex including the frontopolar, dorso-lateral (dlPFC) and ventro-lateral (vlPFC) are activated [BB09]. The DMN along with the various structures in the medial temporal lobe (MTL) including the hippocampus (HF), parahippocampal cortex, perirhinal and entorhinal cortices are also believed to have limited involvement, especially in episodic memory retrieval activities during this phase [BBSS16]. The problem representation requires encoding problem information for which certain

[4] I use the term “agent” to refer to the problem-solver. The term agent is more general than “creature” or “person” or “you" and is intentionally selected to broadly reference humans, animals and artificial agents. I also selectively use the term “solver.”

visual and parietal areas are also involved, although the extent of their involvement is less clear [ASF14, AF14].

Working Memory An important aspect of problem representation is the engagement and use of working memory (WM). The WM allows for the maintenance of relevant problem information and description in the mind [GN12]. Research has shown that WM tasks consistently recruit the dlPFC and left inferior frontal cortex (IC) for encoding and manipulating information; dACC for error detection and performance adjustment; and vlPFC and the anterior insula (AI) for retrieving, selecting information, and inhibitory control [FZZ+16, CW14].

Representation While we mostly believe we know which brain regions are functionally influential in problem definition, less is known about how exactly events are represented *within* these regions. One theory for how events are represented in the PFC is the structured event complex theory (SEC), in which components of the event knowledge are represented by increasingly higher-order convergence zones localized within the PFC, akin to the convergence zones (from posterior to anterior) that integrate sensory information in the brain [BKG09]. Under this theory, different zones in the PFC (left vs right, anterior vs posterior, lateral vs medial and dorsal vs ventral) represent different aspects of the information contained in the events (e.g., number of events to be integrated together, the complexity of the event, whether planning and action is needed). Other studies have also suggested the central executive network's (CEN's) role in tasks requiring cognitive flexibility, and functions to switch thinking modes, levels of abstraction of thought and consider multiple concepts simultaneously [MFE+00].

Thus, when the problem is well-structured, problem representation is largely an executive control activity coordinated by the PFC in which problem information from long-term memory populates WM in a potentially structured representation. Once the problem is defined and encoded, planning and execution of a solution can begin.

9.2.2.2 Planning

The central executive network (CEN), particularly the PFC, is largely involved in plan formation and in plan execution. Planning is the process of generating a strategy to advance from the current state to a goal state. This in turn involves retrieving a suitable solution strategy from memory and then coordinating its execution.

Plan Formation The dlPFC supports sequential planning and plan formation, which includes the generation of hypotheses and construction of plan steps [BB09]. Interestingly, the vlPFC and the angular gyrus (AG), implicated in a variety of functions including memory retrieval, are also involved in plan formation [ASF14]. Indeed, the AG together with the regions in the MTL (including the HF) and several other regions form what is known as the "core" network. The core network is believed to be activated when recalling past experiences, imagining fictitious and future events and navigating large-scale spaces [SHM10], all key functions for generating plan hypotheses. A recent study suggests that the AG is critical to both episodic simulation, representation and episodic memory [TMS17]. One possibility for how plans are formulated could involve a dynamic process of retrieving a promising strategy from memory. Research has shown significant interaction between striatal and frontal regions [SB12, HBB+15]. The striatum is believed to play a key role in declarative memory retrieval, and specifically helping retrieve *optimal* (or previously rewarded) memories [SB12]. Relevant to planning and plan formation, Scimeca & Badre have suggested that the striatum plays two important roles: (1) in mapping acquired value/utility to action selection, and thereby helping plan formation, and (2) modulation and re-encoding of actions and other plan parameters. Different types of problems require different sets of specialized knowledge. For example, the knowledge needed to solve mathematical problems might be quite different (albeit overlapping) from the knowledge needed to select appropriate tools in the environment.

Thus far, I have discussed planning and problem representation as being domain-independent, which has allowed me to outline key areas of the PFC, MTL, and other regions relevant to all problem-solving. However, some types of problems

require domain-specific knowledge for which other regions might need to be recruited. For example, when planning for tool-use, the superior parietal lobe (SPL), supramarginal gyrus (SMG), anterior inferior parietal lobe (AIPL), and certain portions of the temporal and occipital lobe involved in visual and spatial integration have been found to be recruited [BWSH14]. It is believed that domain-specific information stored in these regions is recovered and used for planning.

Plan Execution Once a solution plan has been recruited from memory and suitably tuned for the problem on hand, the left-rostral PFC, caudate nucleus (CN) and bilateral posterior parietal cortices (PPC) are responsible for translating the plan into executable form [SLOA12]. The PPC stores and maintains a "mental template" of the executable form. Hemispherical division of labor is particularly relevant in planning. It was shown that when planning to solve a Tower of Hanoi (block moving) problem, the right PFC is involved in plan construction whereas the left PFC is involved in controlling processes necessary to supervise the execution of the plan [NG15]. On a separate note and not the focus of this chapter, plan execution and problem-solving can require the recruitment of affective and motivational processing in order to supply the agent with the resolve to solve problems, and the vmPFC has been found to be involved in coordinating this process [BB09].

9.2.3 Creativity

During the gestalt movement in the 1930s, Maier noted that "most instances of "real" problem solving involves creative thinking." [Mai30]. Maier performed several experiments to study mental fixation and insight problem solving. This close tie between insight and creativity continues to be a recurring theme, one that will be central to the current discussion. If creativity and insight are linked to RWPS as noted by Maier, then it is reasonable to turn to the creativity and insight literature for understanding the role played by the environment. A large portion of the creativity literature has focused on viewing creativity as an internal process, one in which the solver's attention is directed inwards, and towards internal stimuli, to facilitate

the generation of novel ideas and associations in memory [BBSS16]. Focusing on imagination, a number of researchers have looked at blinking, eye fixation, closing eyes and looking "nowhere" behavior and suggested that there is a shift of attention from external to internal stimuli during creative problem solving [SB16]. The idea is that shutting down external stimuli reduces cognitive load and focuses attention internally. Other experiments studying sleep behavior have also noted the beneficial role of internal stimuli in problem solving. The notion of ideas popping into one's consciousness, suddenly, during a shower is highly intuitive for many and researchers have attempted to study this phenomena through the lens of incubation, and unconscious thought that is internally-driven. There have been several theories and counter-theories proposed to account specifically for the cognitive processes underlying incubation [RD14, Gil16], but none of these theories specifically address the role of the external environment.

The neuroscience of creativity has also been extensively studied and I do not focus on an exhaustive literature review in this chapter (a nice review can be found in [Saw11]). From a problem-solving perspective, it has been found that unlike well-structured problems, ill-structured problems activate the right dlPFC. Most of the past work on creativity and creative problem-solving has focused on exploring memory structures and performing internally-directed searches. Creative idea generation has primarily been viewed as internally directed attention [BJB+16, JBN12] and a primary mechanism involved is *divergent thinking*, which is the ability to produce a variety of responses in a given situation [Gui62]. Divergent thinking is generally thought to involve interactions between the DMN, CEN and the salience network [YR14, HNH+16]. One psychological model of creative cognition is the Geneplore model that considers two major phases of generation (memory retrieval and mental synthesis) and exploration (conceptual interpretation and functional inference) [FWS92, BPP+15]. It has been suggested that the associative mode of processing to generate new creative association is supported by the DMN, which includes the medial PFC, posterior cingulate cortex (PCC), tempororparietal juntion (TPJ), MTL and IPC [BBW+14, BBSS16].

That said, the creativity literature is not completely devoid of acknowledging the role of the environment. In fact, it is quite the opposite. Researchers have looked closely at the role played by externally provided "hints" from the time of the early gestalt psychologists and through to present day studies (Ollinger 2016). In addition to studying how hints can help problem solving, researchers have also looked at how directed action can influence subsequent problem solving – e.g., swinging arms prior to solving the two-string puzzle, which requires swinging the string [TL09]. There have also been numerous studies looking at how certain external perceptual cues are correlated with creativity measures. Vohs et al. suggested that untidiness in the environment and the increased number of potential distractions helps with creativity [VRR13]. Certain colors such as blue have been shown to help with creativity and attention to detail [MZ09]. Even environmental illumination, or lack thereof, have been shown to promote creativity [SW13]. However, it is important to note that while these and the substantial body of similar literature show the relationship of the environment to creative problem solving, they do not specifically account for the cognitive processes underlying the RWPS when external stimuli are received.

9.2.4 Insight Problem Solving

Analytical problem solving is believed to involve deliberate and conscious processing that advances step by step, allowing solvers to be able to explain exactly how they solved it. Inability to solve these problems is often associated with lack of required prior knowledge, which if provided, immediately makes the solution tractable. Insight, on the other hand, is believed to involve a "sudden" and unexpected emergence of an obvious solution or strategy sometimes accompanied by an affective "aha!" experience. Solvers find it difficult to consciously explain how they generated a solution in a sequential manner. That said, research has shown that having an "aha!" moment is neither necessary nor sufficient to insight and vice versa [DWÖ16]. Generally, it is believed that insight solvers acquire a full and deep understanding of the problem when they have solved it [CM11]. There has been an active debate in the problem solving community about whether insight is something special. Some have argued

that it is not, and that there are no special or spontaneous processes, but simply a good old-fashioned search of a large problem space [KS90, MOC01, AW06, Fle08]. Others have argued that insight is special and suggested that it is likely a different process [Dun45, Met86, KB14]. This debate led to two theories for insight problem solving. MacGregor et al., proposed the Criterion for Satisfactory Progress Theory (CSPT), which is based on Newell and Simon's original notion of problem solving as being a heuristic search through the problem space [MOC01]. The key aspect of CSPT is that the solver is continually monitoring their progress with some set of criteria. Impasses arise when there is a criterion failure, at which point the solver tries non-maximal but promising states. The representational change theory (RCT) proposed by Ohlsson et al., on the other hand, suggests that impasses occur when the goal state is not reachable from an initial problem representation (which may have been generated through unconscious spreading activation) [Ohl92]. In order to overcome an impasse, the solver needs to restructure the problem representation, which they can do by (1) elaboration (noticing new features of a problem), (2) re-encoding (fixing mistaken or incomplete representations of the problem, and by (3) changing constraints. Changing constraints is believed to involve two sub-processes of constraint relaxation and chunk-decomposition.

The current position is that these two theories do not compete with each other, but instead complement each other by addressing different stages of problem solving: pre- and post-impasse. Along these lines, Ollinger et al. proposed an extended RCT (eRCT) in which revising the search space and using heuristics was suggested as being a dynamic and iterative or recursive process that involves repeated instances of search, impasse and representational change [ÖJK14, ÖFBS17]. Under this theory, a solver first forms a problem representation and begins searching for solutions, presumably using analytical problem solving processes as described earlier. When a solution cannot be found, the solver encounters an impasse, at which point the solver must restructure or change the problem representation and once again search for a solution. The model combines both analytical problem solving (through heuristic searches, hill climbing and progress monitoring), and creative mechanisms

of constraint relaxation and chunk decomposition to enable restructuring.

Ollinger's model appears to comprehensively account for both analytical and insight problem solving and, therefore, could be a strong candidate to model RWPS. However, while compelling, it is nevertheless an insufficient model of RWPS for many reasons, of which two are particularly significant for the current chapter. First, the model does not explicitly address mechanisms by which external stimuli might be assimilated. Second, the model is not sufficiently flexible to account for other events (beyond impasse) occurring during problem solving, such as distraction, mind-wandering and the like.

So, where does this leave us? I have shown the interplay between problem solving, creativity and insight. In particular, using Ollinger's proposal, I have suggested (maybe not quite explicitly up until now) that RWPS involves some degree of analytical problem solving as well as the post-impasse more creative modes of problem restructuring. I have also suggested that this model might need to be extended for RWPS along two dimensions. First, events such as impasses might just be an instance of a larger class of events that intervene during problem solving. Thus, there needs to be an accounting of the cognitive mechanisms that are potentially influenced by impasses and these other intervening events. It is possible that these sorts of events are crucial and trigger a switch in attentional focus, which in turn facilitates switching between different problem solving modes. Second, we need to consider when and how externally-triggered stimuli from the solver's environment can influence the problem solving process. I detail three different mechanisms by which external knowledge might influence problem solving. I address each of these ideas in more detail in the next two sections.

9.3 Event-Triggered Mode Switching During Problem-Solving

9.3.1 Impasse

When solving certain types of problems, the agent might encounter an impasse, i.e., some block in its ability to solve the problem [SRE+17]. The impasse may arise because the problem may have been ill-defined to begin with causing incomplete and unduly constrained representations to have been formed. Alternatively, impasses can occur when suitable solution strategies cannot be retrieved from memory or fail on execution. In certain instances, the solution strategies may not exist and may need to be generated from scratch. Regardless of the reason, an impasse is an interruption in the problem solving process; one that was running conflict-free up until the point when a seemingly unresolvable issue or an error in the predicted solution path was encountered. Seen as a conflict encountered in the problem-solving process it activates the anterior cingulate cortex (ACC). It is believed that the ACC not only helps detect the conflict, but also switches modes from one of "exploitation" (planning) to "exploration" (search) [QRP08, THPS12], and monitors progress during resolution [CM11]. Some mode switching duties are also found to be shared with the AI (the ACC's partner in the salience network), however, it is unclear exactly the extent of this function-sharing.

Even though it is debatable whether impasses are a necessary component of insight, they are still important as they provide a starting point for creativity [SRE+17]. Indeed, it is possible that around the moment of impasse, the AI and ACC together, as part of the salience network play a crucial role in switching thought modes from analytical planning mode to creative search and discovery mode. In the latter mode, various creative mechanisms might be activated allowing for a solution plan to emerge. Sowden et al. and many others have suggested that the salience network is potentially a candidate neurobiological mechanism for shifting between thinking processes, more generally [SPG15]. When discussing various dual-process

models as they relate to creative cognition, Sowden et al. have even noted that the ACC activation could be a useful marker to identify shifting as participants work creative problems.

9.3.2 Defocused Attention

As noted earlier, in the presence of an impasse there is a shift from an exploitative (analytical) thinking mode to an exploratory (creative) thinking mode. This shift impacts several networks including, for example, the attention network. It is believed attention can switch between a focused mode and a defocused mode. Focused attention facilitates analytic thought by constraining activation such that items are considered in a compact form that is amenable to complex mental operations. In the defocused mode, agents expand their attention allowing new associations to be considered. Sowden et al. (2015) note that the mechanism responsible for adjustments in cognitive control may be linked to the mechanisms responsible for attentional focus. The generally agreed position is that during generative and creative thinking, unconscious cognitive processes activated through defocused attention are more prevalent, whereas during analytical thinking, controlled cognition activated by focused attention becomes more prevalent [SPG15, Kau11].

Defocused attention allows agents to not only process different aspects of a situation, but to also activate additional neural structures in long term memory and find new associations [YR14, Men76]. It is believed that cognitive material attended to and cued by positive affective state results in defocused attention, allowing for more complex cognitive contexts and therefore a greater range of interpretation and integration of information [IDN87]. High attentional levels are commonly considered a typical feature of highly creative subjects [SRE+17].

9.4 Role of the Environment

In much of the past work the focus has been on treating creativity as largely an internal process engaging the DMN to assist in making novel connections in memory.

The suggestion has been that the "individual needs to suppress external stimuli and concentrate on the inner creative process during idea generation" [HNH+16]. These ideas can then function as seeds for testing and problem-solving. While true of many creative acts, this characterization does not capture how creative ideas arise in the solutions to many real-world creative problems. In these types of problems, the agent is functioning and interacting with its environment before, during and after problem-solving. It is natural then to expect that stimuli from the environment might play a role in problem-solving. More specifically, it can be expected that through passive and active involvement with the environment, the agent is (1) able to trigger an unrelated, but potentially useful memory relevant for problem-solving, (2) make novel connections between two events in memory with the environmental cue serving as the missing link, and (3) incorporate completely novel information from events occuring in the environment directly into the problem-solving process. I explore potential neural mechanisms for these three types of environmentally informed creative cognition, which I hypothesize are enabled by defocused attention.

9.4.1 Partial Cues trigger Relevant Memories through Context-Shifting

I have previously discussed the interaction between the MTL and PFC in helping select task-relevant and critical memories for problem-solving. It is well-known that pattern completion is an important function of the MTL and one that enables memory retrieval. Complementary Learning Systems Theory (CLS) and its recently updated version suggest that the MTL and related structures support initial storage as well as retrieval of item and context-specific information [KHM16]. According to CLS theory, the dentate gyrus (DG) and the CA3 regions of the hippocampal formation (HF) are critical to selecting neural activity patterns that correspond to particular experiences [KHM16]. These patterns might be distinct even if experiences are similar and are stabilized through increases in connection strengths between the DG and CA3. Crucially, because of the connection strengths, reactivation of part of a pattern can activate the rest of it (i.e., pattern completion). [KHM16] have

further noted that if consistent with existing knowledge, these new experiences can be quickly replayed and interleaved into structured representations that form part of the semantic memory.

Cues in the environment provided by these experiences hold partial information about past stimuli or events and this partial information converges in the MTL. CLS accounts for how these cues might serve to reactivate partial patterns, thereby triggering pattern completion. When attention is defocused I hypothesize that (1) previously unnoticed partial cues are considered, and (2) previously noticed partial cues are decomposed to produce previously unnoticed sub-cues, which in turn are considered. Zabelina et al. (2016) have shown that real-world creativity and creative achievement is associated with "leaky attention," i.e., attention that allows for irrelevant information to be noticed. In two experiments they systematically explored the relationship between two notions of creativity - divergent thinking and real-world creative achievement - and the use of attention. They found that attentional use is associated in different ways for each of the two notions of creativity. While divergent thinking was associated with flexible attention, it does not appear to be leaky. Instead, selective focus and inhibition components of attention were likely facilitating successful performance on divergent thinking tasks. On the other hand, real-world creative achievement was linked to leaky attention. RWPS involves elements of both divergent thinking and of real-world creative achievement, thus I would expect some amount of attentional leaks to be part of the problem solving process.

Thus, it might be the case that a new set of cues or sub-cues "leak" in and activate memories that may not have been previously considered. These cues serve to reactivate a diverse set of patterns that then enable accessing a wide range of memories. Some of these memories are extra-contextual, in that they consider the newly noticed cues in several contexts. For example, when unable to find a screwdriver, we might consider using a coin. It is possible that defocused attention allows us to consider the coin's edge as being a potentially relevant cue that triggers uses for the thin edge outside of its current context in a coin. The new cues (or contexts) may allow new associations to emerge with cues stored in memory, which can

occur during incubation. Objects and contexts are integrated, in memory, automatically into a blended representation and changing contexts disrupts this recognition [Hay07, Gab16]. Cue-triggered context shifting allows an agent to break apart a memory representation, which can then facilitate problem-solving in new ways.

9.4.2 Heuristic Prototyping facilitates novel associations

It has long been the case that many scientific innovations have been inspired by events in nature and the surrounding environment. As noted earlier, Archimedes realized the relationship between the volume of an irregularly shaped object and the volume of water it displaced. This is an example of heuristic prototyping where the problem-solver notices an event in the environment, which then triggers the automatic activation of a heuristic prototype and the formation of novel associations (between the function of the prototype and the problem) which they can then use to solve the problem [LLQ+13]. Although still in its relative infancy, there has been some recent research into the neural basis for heuristic prototyping. Heuristic prototype has generally been defined as an enlightening prototype event with a similar element to the current problem and is often composed of a feature and a function [HCL+13]. For example, in designing a faster and more efficient submarine hull, a heuristic prototype might be a shark's skin, while an unrelated prototype might be a fisheye camera [DHW+13].

Research has shown that activating the feature function of the right heuristic prototype and linking it by way of semantic similarity to the required function of the problem was the key mechanism people used to solve several scientific insight problems [YDL+16]. A key region activated during heuristic prototyping is the dlPFC and it is believed to be generally responsible for encoding the events into memory and may play an important role in selecting and retrieving the matched unsolved technical problem from memory [DHW+13]. It is also believed that the precuneus plays a role in automatic retrieval of heuristic information allowing the heuristic prototype and the problem to combine [LLQ+13]. In addition to semantic processing, certain aspects of visual imagery have also been implicated in heuristic

prototyping leading to the suggestion of the involvement of Brodmann's area BA 19 in the occipital cortex.

There is some degree of overlap between the notions of heuristic prototyping and analogical transfer (the mapping of relations from one domain to another). Analogical transfer is believed to activate regions in the left medial fronto-parietal system (dlPFC and the PPC) [BB09]. I suggest here that analogical reasoning is largely an internally-guided process that is aided by heuristic prototyping which is an externally-guided process. One possible way this could work is if heuristic prototyping mechanisms help locate the relevant memory with which to then subsequently analogize.

9.4.3 Making Physical Inferences to Acquire Novel Information

The agent might also be able to learn novel facts about their environment through passive observation as well as active experimentation. There has been some research into the neural basis for causal reasoning [BB09, OB16], but beyond its generally distributed nature, we do not know too much more. Beyond abstract causal reasoning, some studies looked into the cortical regions that are activated when people watch and predict physical events unfolding in real-time and in the real-world [FMTK16]. It was found that certain regions were associated with representing types of physical concepts, with the left intraparietal sulcus (IPS) and left middle frontal gyrus (MFG) shown to play a role in attributing causality when viewing colliding objects [MJ13]. The parahippocampus (PHC) was associated with linking causal theory to observed data and the TPJ was involved in visualizing movement of objects and actions in space [MJ13].

9.5 Proposed Theory of Creative Problem Solving

I noted earlier that Ollinger's model for insight problem solving, while serving as a good candidate for RWPS, requires extension. In this section, I propose a candidate model that includes some necessary extensions to Ollinger's framework. I begin by

laying out some preliminary notions that underlie the proposed model.

9.5.1 Dual Attentional Modes

I propose that the attention-switching mechanism described earlier is at the heart of RWPS and enables two modes of operation: focused and defocused mode. In the focused mode, the problem representation is more or less fixed, and problem solving proceeds in a focused and goal directed manner through search, planning and execution mechanisms. In the defocused mode, problem solving is not necessarily goal directed, but attempts to generate ideas, driven by both internal and external items.

At first glance, these modes might seem similar to convergent and divergent thinking modes postulated by numerous others to account for creative problem solving. Divergent thinking allows for the generation of new ideas and convergent thinking allows for verification and selection of generated ideas. So, it might seem that focused mode and convergent thinking are similar and likewise divergent and defocused mode. They are, however, quite different. The modes relate less to idea generation and verification, and more to the specific mechanisms that are operating with regard to a particular problem at a particular moment in time. Convergent and divergent processes may be occurring during both defocused and focused modes. Some degree of divergent processes may be used to search and identify specific solution strategies in focused mode. Also, there might be some degree of convergent idea verification occuring in defocused mode as candidate items are evaluated for their fit with the problem and goal. Thus, convergent and divergent thinking are two among many mechanisms that are utilized in focused and defocused mode. Each of these two modes has to do with degree of attention placed on a particular problem.

There have been numerous dual-process and dual-systems models of cognition proposed over the years. To address criticisms raised against these models and to unify some of the terminology, Evans & Stanovich proposed a dual-process model comprising Type 1 and Type 2 thought [ES13, SPG15]. Type 1 processes are those that are believed to be autonomous and do not require working memory. Type 2

processes, on the other hand, are believed to require working memory and are cognitively decoupled to prevent real-world representations from becoming confused with mental simulations [SPG15]. While acknowledging various other attributes that are often used to describe dual process models (e.g., fast/slow, associative/rule-based, automatic/controlled), Evans & Stanovich note that these attributes are merely frequent correlates and not defining characteristics of Type 1 or Type 2 processes. The proposed dual attentional modes share some similarities with the Evans & Stanovich Type 1 and 2 models. Specifically, Type 2 processes might occur in focused attentional mode in the proposed model as they typically involve the working memory and certain amount of analytical thought and planning. Similarly, Type 1 processes are likely engaged in defocused attentional mode as there are notions of associative and generative thinking that might be facilitated when attention has been defocused. The crucial difference between the proposed model and other dual-process models is that the dividing line between focused and defocused attentional modes is the degree of openness to internal and external stimuli (by various networks and functional units in the brain) when problem solving. Many dual process models were designed to classify the "type" of thinking process or a form of cognitive processing. In some sense, the "processes" in dual process theories are characterized by the type of mechanism of operation or the type of output they produced. Here, I instead characterize and differentiate the modes of thinking by the receptivity of different functional units in the brain to input during problem solving.

This, however, raises a different question of the relationship between these attentional modes and conscious versus unconscious thinking. It is clear that both the conscious and unconscious are involved in problem solving, as well as in RWPS. Here, I claim that a problem being handled is, at any given point in time, in either a focused mode or in a defocused mode. When in the focused mode, problem solving primarily proceeds in a manner that is available for conscious deliberation. More specifically, problem space elements and representations are tightly managed and plans and strategies are available in the working memory and consciously accessible. There are, however, secondary unconscious operations in the focused modes that

includes targeted memory retrieval and heuristic-based searches. In the defocused mode, the problem is primarily managed in an unconscious way. The problem space elements are broken apart and loosely managed by various mechanisms that do not allow for conscious deliberation. That said, it is possible that some problem parameters remain accessible. For example, it is possible that certain goal information is still maintained consciously. It is also possible that indexes to all the problems being considered by the solver are maintained and available to conscious awareness.

9.5.2 RWPS Model

Returning to Ollinger's model for insight problem solving, it now becomes readily apparent how this model can be modified to incorporate environmental effects as well as generalizing the notion of intervening events beyond that of impasses. I propose a theory for RWPS that begins with the standard analytical problem-solving process (See Figures 9.1 and 9.2).

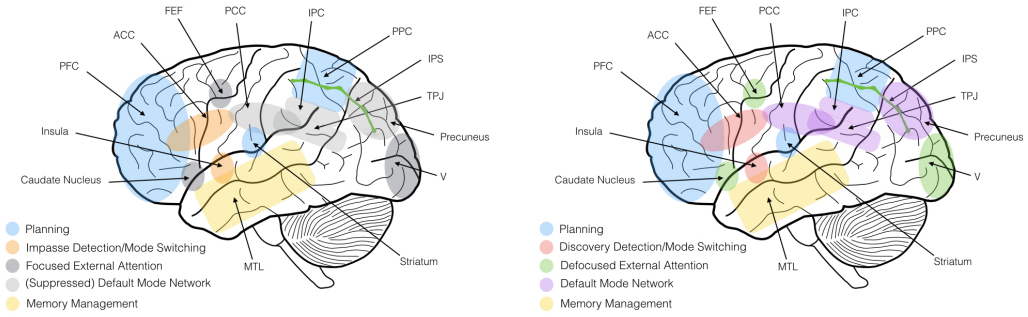

Figure 9.1: Summary of neural activations during focused problem-solving (left) and defocused problem-solving (right). During defocused problem-solving, the salience network (insula and ACC) coordinates the switching of several networks into a defocused attention mode that permits the reception of a more varied set of stimuli and interpretations via both the internally-guided networks (default mode network DMN) and externally guided networks (Attention). PFC: prefrontal cortex, ACC: anterior cingulate cortex, PCC: posterior cingulate cortex, IPC: inferior parietal cortex, PPC: posterior parietal cortex, IPS: intra-parietal sulcus, TPJ: temporoparietal junction, MTL: medial temporal lobe, FEF: frontal eye field.

Focused Problem Solving Mode Initially, both prior knowledge and perceptual entities help guide the creation of problem representations in working memory. Prior

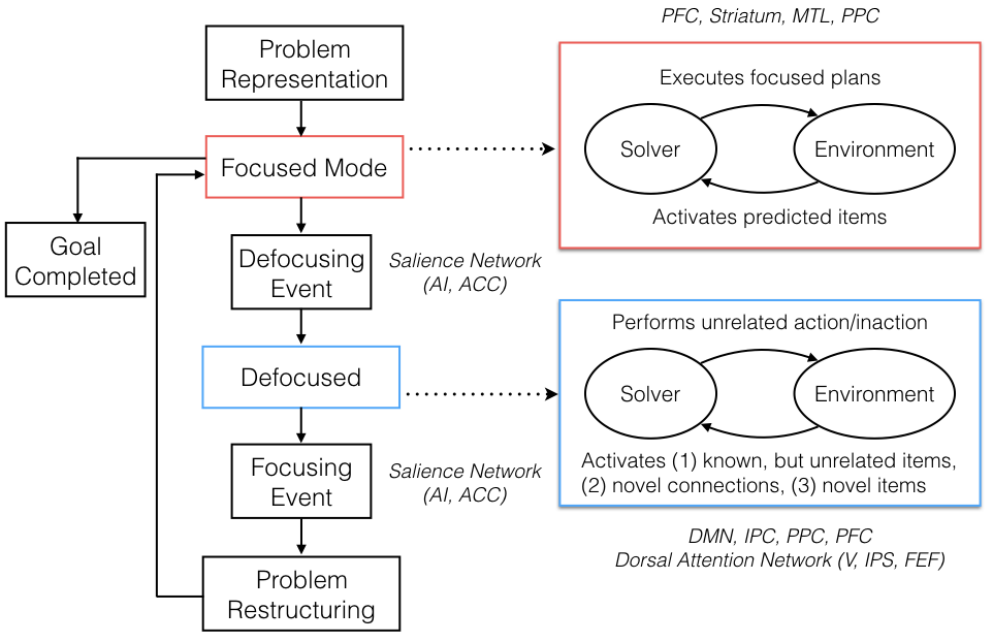

Figure 9.2: Proposed Model for Real World Problem Solving (RWPS). The corresponding neural correlates are shown in italics. During problem-solving, an initial problem representation is formed based on prior knowledge and available perceptual information. The problem-solving then proceeds in a focused, goal-directed mode until the goal is achieved or a defocusing event (e.g., impasse or distraction) occurs. During focused mode operation, the solver interacts with the environment in a directed manner, executing focused plans and allowing for predicted items to be activated by the environment. When a defocusing event occurs, the problem-solving then switches into a defocused mode until a focusing event (e.g., discovery) occurs. In defocused mode, the solver performs actions unrelated to the problem (or is inactive) and is receptive to a set of environmental triggers that activate novel aspects using the three mechanisms discussed in this chapter. When a focusing event occurs, the diffused problem elements cohere into a restructured representation and problem-solving returns into a focused mode.

optimal or rewarding solution strategies are obtained from LTM and encoded in the working memory as well. This process is largely analytical and the solver interacts with their environment through focused plan or idea execution, targeted observation of prescribed entities, and estimating prediction error of these known entities. More specifically, when a problem is presented, the problem representations are activated and populated into working memory in the PFC, possibly in structured represen-

tations along convergence zones. The PFC along with the Striatum and the MTL together attempt at retrieving an optimal or previously rewarded solution strategy from long term memory. If successfully retrieved, the solution strategy is encoded into the PPC as a mental template, which then guides relevant motor control regions to execute the plan.

Defocusing Event-Triggered Mode Switching The search and solve strategy then proceeds analytically until a "defocusing event" is encountered. The salience network (AI and ACC) monitor for conflicts and attempt to detect any such events in the problem-solving process. As long as no conflicts are detected, the salience network focuses on recruiting networks to achieve goals and suppresses the DMN [BBSS16]. If the plan execution or retrieval of the solution strategy fails, then a defocusing event is detected and the salience network performs mode switching. The salience network dynamically switches from the focused problem-solving mode to a defocused problem-solving mode [Men15]. Ollinger's current model does not account for other defocusing events beyond an impasse, but it is not inconceivable that there could be other such events triggered by external stimuli (e.g., distraction or an affective event) or by internal stimuli (e.g., mind wandering).

Defocused Problem Solving Mode In defocused mode, the problem is operated on by mechanisms that allow for the generation and testing of novel ideas. Several large-scale brain networks are recruited to explore and generate new ideas. The search for novel ideas is facilitated by generally defocused attention, which in turn allows for creative idea generation from both internal as well as external sources. The salience network switches operations from defocused event detection to focused event or discovery detection, whereby for example, environmental events or ideas that are deemed interesting can be detected. During this idea exploration phase, internally, the DMN is no longer suppressed and attempts to generate new ideas for problem-solving. It is known that the IPC is involved in the generation of new ideas [BJF+14] and together with the PPC in coupling different information to-

gether [SLOA12, Sim08]. Beaty et al. (2016) have proposed that even this internal idea-generation process can be goal directed, thereby allowing for a closer working relationship between the CEN and the DMN. They point to neuroimaging evidence that supports the possibility that the executive control network (comprising the lateral prefrontal and inferior parietal regions) can constrain and direct the DMN in its process of generating ideas to meet task-specific goals via top down monitoring and executive control [BBSS16]. The control network is believed to maintain an "internal train of thought" by keeping the task goal activated, thereby allowing for strategic and goal-congruent searches for ideas. Moreover, they suggest that the extent of CEN involvement in the DMN idea-generation may depend on the extent to which the creative task is constrained. In the RWPS setting, I would suspect that the internal search for creative solutions is not entirely unconstrained, even in the defocused mode. Instead, the solver is working on a specified problem and thus, must maintain the problem-thread while searching for solutions. Moreover, self-generated ideas must be evaluated against the problem parameters and thereby might need some top-down processing. This would suggest that in such circumstances, we would expect to see an increased involvement of the CEN in constraining the DMN.

On the external front, several mechanisms are operating in this defocused mode. Of particular note are the dorsal attention network, composed of the visual cortex (V), IPS and the frontal eye field (FEF) along with the precuneus and the caudate nucleus allow for partial cues to be considered, as was discussed in Section 9.4.1. The MTL receives synthesized cue and contextual information and populates the WM in the PFC with a potentially expanded set of information that might be relevant for problem-solving. The precuneus, dlPFC and PPC together trigger the activation and use of a heuristic prototype based on an event in the environment. The caudate nucleus facilitates information routing between the PFC and PPC and is involved in learning and skill acquisition.

Focusing Event-Triggered Mode Switching The problem's life in this defocused mode continues until a focusing event occurs, which could be triggered by

either external (e.g., notification of impending deadline, discovery of a novel property in the environment) or internal items (e.g., goal completion, discovery of novel association or updated relevancy of a previously irrelevant item). As noted earlier, an internal train of thought may be maintained that facilitates top-down evaluation of ideas and tracking of these triggers [BBSS16]. The salience network switches various networks back to the focused problem-solving mode, but not without the potential for problem restructuring. As noted earlier, problem space elements are maintained somewhat loosely in the defocused mode. Thus, upon a focusing event, a set or subset of these elements cohere into a tight (restructured) representation suitable for focused mode problem solving. The process then repeats itself until the goal has been achieved.

9.5.3 Model Predictions

Single-Mode Operation The proposed RWPS model provides several interesting hypotheses, which I discuss next. First, the model assumes that any given problem being worked on is in one mode or another, but not both. Thus, the model predicts that there cannot be focused plan execution on a problem that is in defocused mode. The corollary prediction is that novel perceptual cues (as those discussed in Section 9.4) cannot help the solver when in focused mode. The corollary prediction, presumably has some support from the inattentional blindness literature. Inattentional blindness is when perceptual cues are not noticed during a task (e.g., counting the number of basketball passes between several people, but not noticing a gorilla in the scene) [SC99]. It is possible that during focused problem solving, that external and internally generated novel ideas are simply not considered for problem solving. I am not claiming that these perceptual cues are always ignored, but that they are not considered within the problem. Sometimes external cues (like distracting occurrences) can serve as defocusing events, but the model predicts that the actual content of these cues are not themselves useful for solving the specific problem at hand.

When comparing dual-process models Sowden et al. (2015) discuss shifting

from one type of thinking to another and explore how this shift relates to creativity. In this regard, they weigh the pros and cons of serial versus parallel shifts. In dual-process models that suggest serial shifts, it is necessary to disengage one type of thought prior to engaging the other or to shift along a continuum. Whereas in models that suggest parallel shifts, each of the thinking types can operate in parallel. Per this construction, the proposed RWPS model is serial, however, not quite in the same sense. As noted earlier, the RWPS model is not a dual-process model in the same sense as other dual process models. Instead, here, the thrust is on when the brain is receptive or not receptive to certain kinds of internal and external stimuli that can influence problem solving. Thus, while the modes may be serial with respect to a certain problem, it does not preclude the possibility of serial and parallel thinking processes that might be involved within these modes.

Event-driven Transitions The model requires an event (defocusing or focusing) to transition from one mode to another. After all why else would a problem that is successfully being resolved in the focused mode (toward completion) need to necessarily be transferred to defocused mode? These events are interpreted as conflicts in the brain and therefore the mode-switching is enabled by the saliency network and the ACC. Thus, the model predicts that there can be no transition from one mode to another without an event. This is a bit circular, as an event is really what triggers the transition in the first place. But, here I am suggesting that an external or internal cue triggered event is what drives the transition, and that transitions cannot happen organically without such an event. In some sense, the argument is that the transition is discontinuous, rather than a smooth one. Mind-wandering is a good example of when we might drift into defocused mode, which I suggest is an example of an internally driven event caused by an alternative thought that takes attention away from the problem.

A model assumption underlying RWPS is that events such as impasses have an effect similar to other events such as distraction or mind wandering. Thus, it is crucial to be able to establish that there exists a class of such events and they have

a shared effect on RWPS, which is to switch attentional modes.

Focused Mode Completion The model also predicts that problems cannot be solved (i.e., completed) within the defocused mode. A problem can be considered solved when a goal is reached. However, if a goal is reached and a problem is completed in the defocused mode, then there must have not been any converging event or coherence of problem elements. While it is possible that the solver arbitrarily arrived at the goal in a diffused problem space and without conscious awareness of completing the task or even any converging event or problem recompiling, it appears somewhat unlikely. It is true that there are many tasks that we complete without actively thinking about it. We do not think about what foot to place in front of another while walking, but this is not an instance of problem solving. Instead, this is an instance of unconscious task completion.

Restructuring required The model predicts that a problem cannot return to a focused mode without some amount of restructuring. That is, once defocused, the problem is essentially never the same again. The problem elements begin interacting with other internally and externally-generated items, which in turn become absorbed into the problem representation. This prediction can potentially be tested by establishing some preliminary knowledge, and then showing one group of subjects the same knowledge as before, while showing the another group of subjects different stimuli. If the model's predictions hold, the problem representation will be restructured in some way for both groups.

There are numerous other such predictions, which are beyond the scope of this chapter. One of the biggest challenges then becomes evaluating the model to set up suitable experiments aimed at testing the predictions and falsifying the theory, which I address next.

9.6 Experimental Challenges and Paradigms

One of challenges in evaluating the RWPS is that real world factors cannot realistically be accounted for and sufficiently controlled within a laboratory environment. So, how can one controllably test the various predictions and model assumptions of "real world" problem solving, especially given that by definition RWPS involves the external environment and unconscious processing? At the expense of ecological validity, much of insight problem solving research has employed an experimental paradigm that involves providing participants single instances of suitably difficult problems as stimuli and observing various physiological, neurological and behavioral measures. In addition, through verbal protocols, experimenters have been able to capture subjective accounts and problem solving processes that are available to the participants' conscious. These experiments have been made more sophisticated through the use of timed-hints and/or distractions. One challenge with this paradigm has been the selection of a suitable set of appropriately difficult problems. The classic insight problems (e.g., Nine-dot, eight-coin) can be quite difficult, requiring complicated problem solving processes, and also might not generalize to other problems or real world problems. Some in the insight research community have moved in the direction of verbal tasks (e.g., riddles, anagrams, matchstick rebus, remote associates tasks, and compound remote associates tasks). Unfortunately, these puzzles, while providing a great degree of controllability and repeatability, are even less realistic. These problems are not entirely congruent with the kinds of problems that humans are solving every day.

The other challenge with insight experiments is the selection of appropriate performance and process tracking measures. Most commonly, insight researchers use measures such as time to solution, probability of finding solution, and the like for performance measures. For process tracking, verbal protocols, coded solution attempts, and eye tracking are increasingly common. In neuroscientific studies of insight various neurological measures using functional magnetic resonance imaging (fMRI), electroencephalography (EEGs), transcranial direct current stimulation (tDCS) and

transcranial magnetic stimulation (tMS) are popular and allow for spatially and temporally localizing an insight event.

Thus, the challenge for RWPS is two-fold: (1) selection of stimuli (real world problems) that are generalizable, and (2) selection of measures (or a set of measures) that can capture key aspects of the problem solving process. Unfortunately, these two challenges are somewhat at odds with each other. While fMRI and various neuroscientific measures can capture the problem solving process in real time, it is practically difficult to provide participants a realistic scenario while they are laying flat on their back in an fMRI machine and allowed to move nothing more than a finger. To begin addressing this conundrum, I suggest returning to object manipulation problems (not all that different from those originally introduced by Maier and Duncker nearly a century ago), but using modern computing and user-interface technologies.

One pseudo-realistic approach is to generate challenging object manipulation problems in Virtual Reality (VR). VR has been used to describe 3-D environment displays that allows participants to interact with artificially projected, but experientially realistic scenarios. It has been suggested that virtual environments (VE) invoke the same cognitive modules as real equivalent environmental experience [For10]. Crucially, since VE's can be scaled and designed as desired, they provide a unique opportunity to study pseudo-RWPS. However, a VR-based research approach has its limitations, one of which is that it is nearly impossible to track participant progress through a virtual problem using popular neuroscientific measures such as fMRI because of the limited mobility of connected participants.

Most of the studies cited in this chapter utilized an fMRI-based approach in conjunction with a verbal or visual task involving problem-solving or creative thinking. Very few, if any, studies involved the use physical manipulation, and those physical manipulations were restricted to limited finger movements. Thus, another pseudo-realistic approach is allowing subjects to teleoperate robotic arms and legs from inside the fMRI machine. This paradigm has seen limited usage in psychology and robotics, in studies focused on Human-Robot interaction [LJGDR15]. It

could be an invaluable tool in studying real-time dynamic problem-solving through the control of a robotic arm. In this paradigm a problem solving task involving physical manipulation is presented to the subject via the cameras of a robot. The subject (in an fMRI) can push buttons to operate the robot and interact with its environment. While the subjects are not themselves moving, they can still manipulate objects in the real world. What makes this paradigm all the more interesting is that the subject's manipulation-capabilities can be systematically controlled. Thus, for a particular problem, different robotic perceptual and manipulation capabilities can be exposed, allowing researchers to study solver-problem dynamics in a new way. For example, even simple manipulation problems (e.g., re-arranging and stacking blocks on a table) can be turned into challenging problems when the robotic movements are restricted. Here, the problem space restrictions are imposed not necessarily on the underlying problem, but on the solver's own capabilities. Problems of this nature, given their simple structure, may enable studying everyday practical creativity without the burden of devising complex creative puzzles. Crucial to note, both these pseudo-realistic paradigms proposed demonstrate a tight interplay between the solver's own capabilities and their environment.

9.7 Conclusion

While the neural basis for problem-solving, creativity and insight have been studied extensively in the past, there is still a lack of understanding of the role of the environment in informing the problem-solving process. Current research has primarily focused on internally-guided mental processes for idea generation and evaluation. However, the type of real world problem-solving (RWPS) that is often considered a hallmark of human intelligence has involved both a dynamic interaction with the environment and the ability to handle intervening and interrupting events. In this chapter, I have attempted to synthesize the literature into a unified theory of RWPS, with a specific focus on ways in which the environment can help problem-solve and the key neural networks involved in processing and utilizing relevant and useful en-

vironmental information. Understanding the neural basis for RWPS will allow us to be better situated to solve difficult problems. Moreover, for researchers in computer science and artificial intelligence, clues into the neural underpinnings of the computations taking place during creative RWPS, can inform the design of the next generation of helper and exploration robots which need these capabilities in order to be resourceful and resilient in the open-world.

Chapter 10

Formalizing the MacGyver Problem

In the previous chapter, we considered a possible model for human problem solving and suggested that it might inform how we design creative AI systems. In this chapter, we formalize this mathematically in the language of Classical Planning and name them MacGyver Problems, in honor of the widely popular television series from the 1980s that exemplified and celebrated resourcefulness and creativity in high-stakes situations. We argue for creative problem solving to not only be a useful skill for machines but also a useful test of intelligence.

10.1 Introduction

How should we evaluate machine intelligence? This a long-standing problem in AI and robotics. From Alan Turing's original question about whether machines can think [Tur50] to today's plethora of robotics and AI challenges [LDM12b, Fei03, Bod10, BS16b, Coh05, Har91, Rie14] and data sets [JHvdM+17, WBC+15] , the question of what makes a suitable test is still open, relevant, and crucial to judging progress in AI while guiding its research and development.

The crux of this question is a choice about what we should measure. The Turing Test focused on natural language interactions; its progenies have since expanded

to include vision, planning, game playing, localization and mapping and many others. Research in various AI subfields is often guided by these sorts of targeted data sets and challenges. Since research questions within AI subfields are quite rich and complex, we might argue there is no need to pick one behavior, but rather pursue all of them, separately and in parallel. However, psychological and zoological studies in human and animal intelligence suggest the existence of general capabilities that transcend these types of targeted abilities [Ack16].

We introduce the idea that general intelligence is encapsulated in notions of resourcefulness, improvisation, and creative problem solving that we humans use every day and that we celebrate in movies like "The Martian" and television shows like "MacGyver." We thus ask a different question: *can machines improvise when they become stuck on a problem?*

As the first step towards formalizing intuitions of creative problem solving and improvisation, we define a new class - *MacGyver problems* - using the language of classical planning. The idea here is to let an agent begin with a seemingly unsolvable problem and observe how it fares in attempting to find a solution. If the agent can take actions and incorporate relevant knowledge from the environment to make the problem more tractable (or at least computable in some cases), then we might be able to predict the agent's capabilities in solving novel problems in the open world. We present these ideas in the context of an illustrative blocks world variant called *cup world* (shown in Figure 10.1), where we define an initial instance that is unsolvable. The agent must then transform and modify its initial domain representations to learn and incorporate new concepts and actions that let it solve the problem.

In the following pages, we define the MacGyver framework more formally (Section 10.3) and provide some complexity results. In Section 10.4, we return to the cup world domain and outline how a MacGyver problem in cup world could be solved, while noting the underlying challenges in solving these problems, more generally. We then discuss, in Section 10.5, the space of capabilities and component subtasks that an agent might need to solve MacGyver problems. In Section 10.6, we discuss how we might evaluate these agents and how the research community

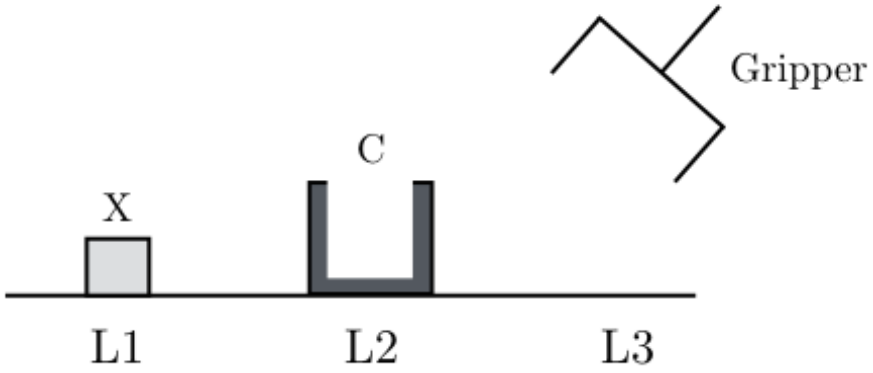

Figure 10.1: Cups world: contains a block X and a cup C. It is easy to see how the block can be moved from location $L1$ to $L3$. But, what if the gripper is not allowed to touch the block? Details in Section 10.4.

can evaluate its progress. Finally, Section 10.7 concludes and discusses avenues for future research. But first we briefly review the history of machine intelligence tests and highlight some of their missing aspects in Section 10.2.

10.2 The Turing Test and its Progeny

Alan Turing (1950) asked whether machines could produce observable behavior (e.g., natural language) that we would say required thought in people. He suggested that if interrogators were unable to tell, after having long free-flowing conversations with a machine whether they were dealing with a machine or a person, then they could conclude that the machine was "thinking". Turing did not actually intend for this to be a test, but rather a prediction of what could be achieved, technologically, in fifty years [CVL13]. Nevertheless, others have since developed tests for machine intelligence that were variations of the so-called Turing Test to address a common criticism that it was easy to deceive the interrogator.

For example, [LDM12b] designed a reading comprehension test, entitled the *Winograd Schema Challenge*, in which the agent is presented a question having some ambiguity in the referent of a pronoun or possessive adjective. The question asks the reader to determine the referent of this ambiguous pronoun or possessive adjective by selecting one of two choices. [Fei03] proposed a variation of the Turing

Test in which a machine can be evaluated by a team of subject matter specialists through natural language conversation. Other proposed tests have attempted to study a machine's ability to produce creative artifacts and solve novel problems [Bod10, BBF01, BS16b, Rie14, Wig06].

Extending capabilities beyond the linguistic and creative ones, [Har91] suggested a *Total Turing Test* (T3) that expanded the range of capabilities to a full set of robotic capacities found in embodied systems. [Sch12] extended the T3 to incorporate species evolution and development over time, proposing the *Truly Total Turing Test* (T4) to test not only individual cognitive systems but whether the candidate cognitive architecture is capable, as a species, of long-term evolutionary achievement.

Finding that the Turing Test and its variants were not helping guide research and development, others proposed a task-based approach. Specific task-based goals were couched as toy problems that were representative of a real-world task [Coh05]. The research communities benefited greatly from this approach and focused their efforts towards specific machine capabilities like object recognition, automatic scheduling and planning, scene understanding, localization and mapping, and even game playing. Many competitions and challenges emerged that tested the machine's performance in applying these capabilities. Some competitions even tested embodiment and robotic capacities that combined multiple tasks. For example, the DARPA *Robotics Challenge* evaluated a robot's ability to conduct remote operations including turning valves, using a tool to break through a concrete panel, opening doors, and remove debris from entryways.

Unfortunately, neither the Turing Test variants nor the task-based challenges captured the intuitive notions of resourcefulness that is at the core of creative problem solving. Creative agents situated in the real world must be able to solve problems with limited resources. This means they must be able to improvise, challenge prior assumptions, generate and test ideas, make new observations, and acquire new knowledge from their environments.

10.3 The MacGyver Framework

Our core proposal is to present an agent with a problem that is unsolvable with its initial knowledge and observe its problem-solving processes to estimate the degree to which it is creative. If the agent can think outside of its current context, take exploratory actions, incorporate relevant environmental cues, and learn knowledge to make the problem tractable (or at least computable), then it has the general ability to solve open-world problems.

This notion of expanding one's current context builds on Boden's ([Bod10]) seminal work on creativity and, specifically, her distinction between *exploratory* and *transformational* creativity. Boden introduced the notion of a conceptual space and proposed that exploratory creativity involves discovering new objects in this space. Transformational creativity involves redefining that conceptual space and producing a paradigm shift. There have been a few attempts to formalize and conceptualize these ideas, including Wiggins ([Wig06]) Creative Systems Framework. He formalized Boden's notion of a conceptual space $\mathscr{C}$ that is a subset of a universe $\mathscr{U}$, that in turn contains every possible concept. A conceptual space, according to Wiggins, is defined by a set of rules $\mathscr{R}$ for constraining the space and a set of rules $\mathscr{T}$ for traversing it. Exploratory creativity occurs when the rule sets are used within a concept space to discover new concepts. Transformational creativity occurs when the conceptual space is redefined by modifying rule sets. Using this abstract formalism, he began the task of characterizing the behavior of creative systems and, importantly, distinguishing between object-level search within a conceptual space and meta-level searches of conceptual spaces.

However, more work was still needed to operationalize these concepts, which, while more formal than Boden's original proposal, lacked depth and connections to work in AI and robotics formalisms. More recently, [CBCH16] built on Wiggins' basic ideas and proposed that the dual searches - exploratory and transformational - can be set up as a hierarchical reinforcement learning problem formalized in terms of Markov decision processes. While this is a promising approach to formalizing these

dual searches, we believe that it is just one piece of a larger puzzle. In response, we propose an approach that unifies several AI research traditions including not only reinforcement learning but also planning, belief revision, and others. We can now start to formalize the MacGyver framework.

10.3.1 Formal Definition of a MacGyver Problem

We define $\mathcal{L}$ to be a first-order language with atoms $p(t_1, \ldots, t_n)$ and their negations $\neg p(t_1, \ldots, t_n)$, where t_i represents terms that can be variables or constants. An atom is grounded if and only if all of its terms are constants. A planning domain in $\mathcal{L}$ can be represented as $\Sigma = (S, A, \gamma)$, where S denotes the set of states, A the set of actions, and γ the transition functions. A planning problem $\mathcal{P} = (\Sigma, s_0, g)$ consists of the domain, the initial state s_0, and the goal state g. A plan π is any sequence of actions. A plan π is a solution to the planning problem if $g \subseteq \gamma(s_0, \pi)$. We also consider the notion of state reachability and the set of all successor states $\hat{\Gamma}(s)$, which defines the set of states reachable from s.

To formalize a MacGyver problem, we define a universe and a world within this universe. The world describes the full set of abilities of an agent and includes those abilities that the agent knows about and those of which it is unaware. We can then define an agent subdomain as representing a proper subset of the world that is within the agent's awareness. A MacGyver problem then becomes a planning problem that is defined in the world, but outside the agent's current subdomain, as depicted in Figure 10.2.

Definition *(Universe).* A universe $\mathbb{U} = (S, A, \gamma)$ is a classical planning domain that represents all aspects of the physical world perceivable and actionable by any and all agents, regardless of capabilities. This includes all allowable states, actions, and transitions in the physical universe.

Definition *(World).* A world $\mathbb{W}^t = (S^t, A^t, \gamma^t)$ is a portion of the Universe $\mathbb{U}$ corresponding to those aspects that are perceivable and actionable by a particular species t of agent. Each agent species $t \in T$ has a particular set of sensors and

actuators allowing agents in that species to perceive a proper subset of states, actions or transition functions. Thus, a world can be defined as

$$\mathbb{W}^t = \{(S^t, A^t, \gamma^t) \mid ((S^t \subseteq S) \vee (A^t \subseteq A) \vee (\gamma^t \subseteq \gamma)) \wedge \neg((S^t = S) \wedge (A^t = A) \wedge (\gamma^t = \gamma))\}.$$

Definition *(Agent Subdomain).* An agent subdomain $\Sigma_i^t = (S_i^t, A_i^t, \gamma_i^t)$ of type t is a planning subdomain that corresponds to the agent's perception and action within its world $\mathbb{W}^t$. In other words, the agent is not fully aware of all of its capabilities at all times, and the agent domain Σ_i^t corresponds to the portion of the world that it is perceiving and acting at time i.

$$\Sigma_i^t = \{(S_i^t, A_i^t, \gamma_i^t) \mid ((S_i^t \subset S^t) \vee (A_i^t \subset A^t) \vee (\gamma_i^t \subset \gamma^t)) \wedge \neg((S_i^t = S^t) \wedge (A_i^t = A^t) \wedge (\gamma_i^t = \gamma^t))\}$$

Definition *(MacGyver Problem).* A MacGyver problem with respect to an agent t is a planning problem in the agent's world $\mathbb{W}_t$ that has a goal state g that is currently unreachable by the agent. Formally, a MacGyver problem $\mathcal{P}_M = (\mathbb{W}^t, s_0, g)$, where

- $s_0 \in S_i^t$ is the initial state of the agent
- g is a set of ground atoms
- $S_g = \{s \in S | g \subseteq s\}$

Where $g \subseteq s', \forall s' \in \hat{\Gamma}_{\mathbb{W}^t}(s_0) \setminus \hat{\Gamma}_{\Sigma_i^t}(s_0)$.

10.3.2 Complexity Results and Importance of Heuristics

It naturally follows that, in the context of a world $\mathbb{W}_t$, the MacGyver problem $\mathcal{P}_M$ is a classical planning task which from the agent's current perspective is unsolvable. We can reformulate the MacGyver problem in terms of language recognition to analyze its complexity. This will clarify the difficulty of the problem, and, importantly, establish the key role that heuristics play when solving MacGyver problems.

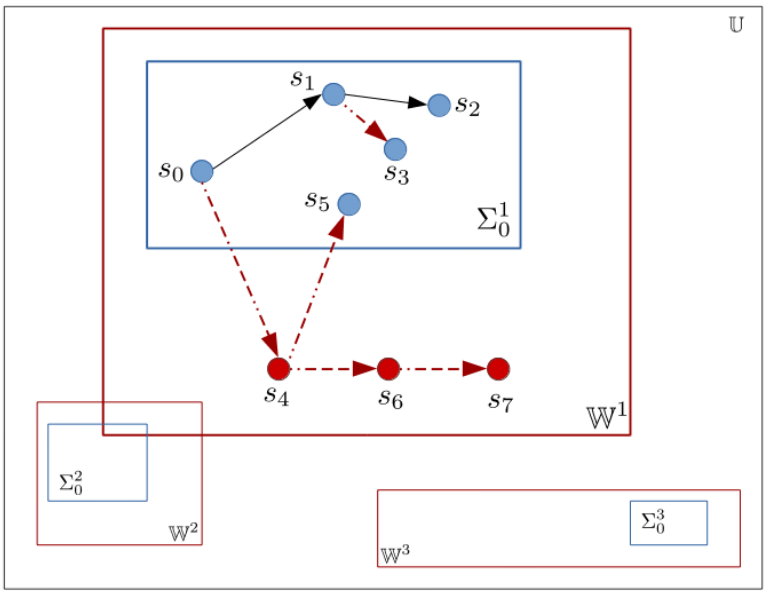

Figure 10.2: Conceptual diagram showing several exemplary MacGyver problems and how they relate to classical planning tasks. Three agents ($t = 1, 2, 3$) are depicted along with their starting domains (Σ_0^t) and their worlds ($\mathbb{W}^t$). Operators and states not in the agent's initial domain are shown in red. Thus, states s_3, s_4, s_5, s_6, s_7 are unreachable from s_0 in domain Σ_0^1, and consequently could be characterized as goal states of MacGyver problems, albeit of different difficulties.

Definition *(MGP-EXISTENCE).* Given a set of statements D of planning problems, let *MGP-EXISTENCE(D)* be the set of all statements $P \in D$ such that P represents a MacGyver problem $\mathcal{P}_M$.

Theorem 10.3.1 *MGP-EXISTENCE is decidable.*

Proof The number of possible states in the agent's subdomain Σ_i^t and the agent's world $\mathbb{W}^t$ are finite. So, it is possible to do a brute-force search to see whether a solution exists in the agent's world but not in the agent's initial domain.

Theorem 10.3.2 *MGP-EXISTENCE is EXPSPACE-complete.*

Proof *(*Membership). An MGP amounts to looking to see if the problem is a solvable problem in the agent-domain. Upon concluding it is not solvable, the problem then becomes one of determining if it is a solvable problem in the world corresponding to the agent's species. Each of these problems are PLAN-EXISTENCE

problems, which are in EXPSPACE for the unrestricted case [GN04]. Thus, MGP-EXISTENCE is in EXPSPACE.

*(*Hardness). We can reduce the classical planning problem $P(\Sigma, s_0, g)$ to an MGP (PLAN-EXISTENCE $\leq_m^p$ MGP-EXISTENCE), by defining a new world $\mathbb{W}$. To define a new world, we extend the classical domain by one state, defining the new state as a goal state, and adding actions and transitions from every state to the new goal state. We also set the agent domain to be the same as the classical planning domain. Now, $P(\Sigma, s_0, g) \in$ PLAN-EXISTENCE iff $P(\mathbb{W}, s_0, g) \in$ MGP-EXISTENCE for agent with domain Σ. Thus, MGP-EXISTENCE is EXPSPACE-*hard*.

Theorems 1 and 2 show that the question of knowing whether a given problem is a MacGyver problem, although computable (in the finite case), is still intractable.[1] Thus, the agent will necessarily need heuristics to explore its own search space. We argue that the role of heuristics in solving these problems is paramount. Indeed, there is no single heuristic in any sense of the word [Lan17] that can guide an agent in all MacGyver problems. Even for a given MacGyver task, it is unlikely that a single heuristic will be sufficient. This means that the agent must use and reason with a family of heuristics as it searches for a solution. This is good news for cognitive systems research that aims to leverage the power of heuristics in problem solving and align it with parallel results in humans, and, conversely, uncover heuristics that people use when solving these problems to better inform agent design. In fact, we might flip how we think about heuristics for MacGyver problems: from a weak method that either produces a suboptimal solution or does not guarantee a solution at all, to a necessary step towards achieving representational change or domain transformation, thereby eliciting satisfactory solutions.

[1]The proofs for these theorems are straightforward applications of complexity-theoretic techniques and have been omitted here for brevity.

10.3.3 MacGyver Solution via Domain Modification

By definition, MacGyver problems require that the initial domain be transformed in some way (e.g., by adding a state, transition function, or action) for the goal state to be reachable. Here, we provide some formal specifications for what a solution strategy might look like for a MacGyver agent, which must keep modifying its domain representation until it can reach the goal.

Definition *(Agent Domain Modification).* A domain modification Σ_j^{t*} involves either a domain extension or contraction (for brevity, we only consider extensions here). A domain extension Σ_j^{t+} of an agent is an Agent-subdomain at time j that is in the agent's world $\mathbb{W}^t$ but not in the agent's subdomain Σ_i^t in the previous time i, such that Σ_i^t precedes Σ_j^t (i.e., $\Sigma_i^t \preceq \Sigma_j^t$). The agent extends its subdomain through sensing and perceiving its environment and its own self. E.g., the agent can extend its domain by making an observation, receiving advice or an instruction or performing introspection. Formally,

$$\Sigma_j^{t+} = \{(S_j^{t+}, A_j^{t+}, \gamma_j^{t+}) \mid (S_j^{t+} \subset S^t \setminus S_i^t) \vee (A_j^{t+} \subset A^t \setminus A_i^t) \vee (\gamma_j^{t+} \subset \gamma^t \setminus \gamma_i^t)\}.$$

The agent subdomain that results from a domain extension is $\Sigma_j^t = \Sigma_i^t \cup \Sigma_j^{t+}$. A domain modification set $\Delta_{\Sigma_i^t} = \{\Sigma_1^{t*}, \Sigma_2^{t*}, \ldots, \Sigma_n^{t*}\}$ is a set of n domain modifications on subdomain Σ_i^t. Let Σ_Δ^t be the subdomain resulting from performing the modifications in $\Delta_{\Sigma_i^t}$ on Σ_i^t.

Definition *(Strategy and Domain-Modifying Strategy).* A *strategy* is a tuple $\omega = (\pi, \Delta)$ of a plan π and a set Δ of domain modifications. A domain-modifying strategy ω^C involves at least one domain modification, i.e., $\Delta \neq \emptyset$.

Definition *(Context).* A *context* is a tuple $\mathbb{C}_i = (\Sigma_i, s_i)$ that represents the agent's subdomain and state at time i.

We are now ready to define an insightful strategy as a set of actions and domain modifications that the agent must perform for the problem's goal state to

be reachable.

Definition *(Insightful Strategy).* Let $\mathbb{C}_i = (\Sigma_i^t, s_0)$ be the agent's current context. Let $\mathcal{P}_M = (\mathbb{W}^t, s_0, g)$ be a MacGyver problem for the agent in this context. An insightful strategy is a domain-modifying strategy $\omega^I = (\pi^I, \Delta^I)$ that, when applied in $\mathbb{C}_i$, results in a context $\mathbb{C}_j = (\Sigma_j^t, s_j)$, where $\Sigma_j^t = \Sigma_{\Delta^I}^t$ such that $g \subseteq s', \forall s' \in \hat{\Gamma}_{\Sigma_j^t}(s_j)$.

Formalizing the insightful strategy in this way is analogous to the moment of insight reached when a problem becomes tractable (or in our definition computable) or when a solution plan becomes feasible. Specifically, solving a problem involves creative exploration, domain extensions, and domain contractions until the agent has the information it needs within its subdomain to solve it as a classical planning task, and it does not need any further domain extensions.

Note, however, the definitions stated here do not require a multitude of domain transformations. The agent might achieve a flash of insight about the appropriate representation in a single transformation. Alternatively, the agent may work through several minor transformations. So the definitions here are meant to capture creative problem solving more generally, encompassing both single-step and multi-step insights.

Moreover, the definition of an insightful strategy does not require using a single strategy for all MacGyver problems, as no such universal strategy exists. Instead, definition 9 merely serves as a formalism for capturing notions of context and representational change that apply to any solution. That is, the definition is not tied to a particular algorithm or solution plan, and accordingly does not speak to the quality of the solution reached. An insightful strategy is just meant to move the agent to a state from which the goal state is reachable. Achieving the goal state might still be intractable, as is any classical planning problem.

10.3.4 Connections to Insight Problem Solving in Humans

The MacGyver problem formulation proposed in this essay draws substantial inspiration from decades-old lines of psychological research on insight problem solving. There are numerous theories of insight and its links to creativity, one of the earliest being Wallas' ([Wal26]) four-stage framework of preparation, incubation, illumination, and verification. Many subsequent psychological theories have addressed what solvers were doing during each of these stages. What these theories shared was that insight problem solvers encountered an impasse and later had an insight that let them solve the task [Ohl92]. In this essay, we consider the problem-space approach to insight problem solving, in which impasses occur due to an incorrectly chosen problem representation, such as wrong propositions or operators) [KS90, Ohl92]. Impasses can be broken by revising the problem representation, which in turn can be viewed as a meta-level search over the space of representations. An alternative approach to thinking about insight is to consider it not as a search through a problem space and then a meta-level space, but as a memory retrieval task of indexing, retrieving, and applying an analogous knowledge structure [LJ88].

Notwithstanding their differences, all the theories agree about the importance of cues and heuristics in the search (or as appropriate, analogical retrieval) process. These cues and heuristics help the solver consider previously ignored knowledge structures, propositions, relations, and operators, as well as consider new ones. The heuristics themselves may be strategies for navigating through or modifying the knowledge structures (e.g., problem representations) that contain the problem or solution. In addition, some heuristic strategies might actively or passively participate with the surrounding environment and observe what happens [Sar18a]. The purpose of these cues and heuristics is to arrive at a representation that allows a full or partial insight. We preview a smattering of such heuristics in the next section, where we provide a blueprint for solving a MacGyver problem in the context of the cup world domain introduced in Section 10.1. Overall, we believe the connections between the MacGyver problem solving and insight problem solving in humans to be deeply

interesting and worthy of further research.

10.4 A Conceptual Blueprint for Solving Cup World

We are now ready to operationalize the MacGyver framework in the cup world domain, outline the domain transformations needed to solve this problem, and discuss some domain-modifying strategies. This will highlight the sorts of computational, perceptual, and action capabilities needed to execute the appropriate strategy, assimilate the necessary knowledge, and make the needed domain transformations.

The cup world domain consists of a block and cup on a table, as shown in Figure 10.1. Consider a start domain Σ_0^t for an agent t with a gripper. Let us say the agent has the ability to perceive and reason about relational aspects of its environment (e.g., whether an object is visible, whether it has been touched, and spatial relationships between objects). The domain is discrete and finite, with actions being represented symbolically. Let us also assume that the agent is a pick-and-place robot with a single operator for picking and placing objects; this *pick-and-place* action has an associated script composed of several sub-actions ($reach\text{-}for, grasp, lift, move\text{-}over$, $set\text{-}down, release$), that form the set of primitive actions.

The agent is primarily tasked with picking and placing objects, and accordingly has chunked subactions into a single *pick-and-place* macro-operator that increases planning efficiency. That said, the primitive actions can be directly coupled to robotic actuators with precise continuous values. For example, the *lift* action can be parametrized to 3D point clouds that represent objects in space and real-valued vertical displacements. Later in this section, we comment on extensions to infinite and continuous space in which new primitive actions can be generated from parameterizations of actuator controls. Thus, we assume here that the robot can process 3D point cloud information from the environment and determine whether an object is a cup or a block, and whether it is upright.

In Σ_0^t, we can solve traditional planning problems such as moving the block x from l_1 to l_3, with the plan $\pi = \{pick\text{-}and\text{-}place(x, l_1, l_3)\}$. We can define a

MacGyver problem, $\mathcal{P}_M$, for this agent t given a start domain Σ_0^t by requiring that the block x be moved to l_3 **without touching it**, that is $touched(x) = False$. To handle this problem, the agent must extend its start domain by using heuristics and cues. Although there may be many different ways to solve this MacGyver problem, let us consider one particular approach and analyze more closely what sorts of heuristics and cues might be activated by an agent conforming to it. Consider a set of two domain extensions $\Delta_{\Sigma_0^t} = \{\Sigma_1^{t+}, \Sigma_2^{t+}\}$. The first extension Σ_1^{t+} includes a *nudge* action and the second domain extension Σ_2^{t+} includes several new actions and the fluent *enclosed*.

We can decompose the *pick-and-place* action into its primitive actions (*Heuristic 1:* decompose chunks) and then attempt to replan with these primitive operators. However, this will fail, as any action that requires grasping the block will trigger the *touched* event. Thus, the agent may instead attempt to solve a simpler problem of moving the block without picking it up (*Heuristic 2:* reformulate goals and define simpler goals). [LPB+16] have reported an architecture for hierarchical problem solving that searches through a space of problem decompositions, which might be a fruitful approach here. The *lift* action triggers the *pickedup* event, so the agent must avoid this action. This means that the *lift* action's postcondition of *holding* must be relaxed as a requirement from the *move-across* action (*Heuristic 3:* relax constraints and preconditions). The agent then defines a new action, *nudge*, which is essentially the *move-across* action without the *holding* requirement. Upon performing this action in the environment, the agent notices that the cup c can be moved from one location to another without picking it up (*Heuristic 4:* notice invariants). This is a major milestone in the agent's problem solving. It can now solve the intermediate problem (originally a MacGyver problem, as well) by following the plan

$$\pi = \langle reach\text{-}for(c), grasp(c), nudge(c, l_3), release(c), reach\text{-}for(x), grasp(x),$$
$$nudge(x, l_2), release(x), pick\text{-}and\text{-}place(c, l_3, l_1), reach\text{-}for(x), nudge(x, l_3)\rangle.$$

The agent might not know it yet, but it is much closer to solving the more difficult "not-touched" problem. Next, the agent might attempt to relax constraints on other primitive actions and discover that it can pick up, flip, and set down the cup on top of the block, as *set-down* otherwise requires a clear location. In doing so, it might notice something unexpected, namely that the block disappears (*Heuristic 5:* notice anomaly). The agent can then hypothesize a new relation called $P(x, c, l_1)$ to account for this behavior and generate a new action, *cover*, to capture the dynamics (*Heuristic 6:* hypothesize predicates). After experimenting with this process, and through language-based human assistance, it might rename the predicate P to be *enclosed*.

Finally, with additional experimentation, the agent must learn a *scoot* action, which is a variant of *nudge* with added knowledge that there will be co-movement of the block and cup if the cup encloses the block. The purpose of this example is not to demonstrate a generalizable solution to a MacGyver problem, which no one has achieved to date. However, it does provide a rough sense of the types of knowledge and heuristics that could support creative problem solving in this domain.

Thus far, we have treated perceptual capabilities and primitive actions as symbols in cup world. However, in real-world settings, these primitive actions will be grounded in robotic actions that are parametrized as points in continuous 3D space, with numeric color and depth values. We can extend our formulation to infinite MacGyver problems that, informally, require the agent to synthesize new symbols from real-valued grounded actions and percepts (e.g., learning to detect *maroon* when it has a color detector and knowledge of the symbol *red*). We can pursue such questions in conjunction with research on integrating task and motion planning, and on learning symbolic operators from continuous spaces [MZPS12, KKLP14, KLP13, SFR+14, GLPK17].

10.5 Agent Capabilities and Subtasks

As we have seen, solving even simple MacGyver problems can be a challenging endeavor, requiring the ability to track and maintain cues and heuristics while remaining cognizant of the current state of problem solving. Nevertheless, we are optimistic that research progress can be made in this regard if we can leverage results from relevant subfields in AI and cognitive systems. Below is short, nonexhaustive list of subtasks that MacGyver problems appear to require:

1. Impasse Detection. As noted earlier, determining the existence of a MacGyver problem is intractable. Thus, the agent must use heuristics in an attempt to solve a MacGyver problem as a traditional planning task. It must also be equipped to detect unsolvability or at least intractability [BJS13, LMG16]. Recent efforts by the planning community have started to address this issue through the "First Unsolvability IPC" at ICAPS 2016.

2. Domain Transformation and Problem Restructuring. Chunk decomposition and constraint relaxation have been well studied in psychology [Kno09]. The extensive literature in plan task revision has formally studied the effects of changing states, goal, and operators [HdMdBW14, GKE+10].

3. Experimentation. The agent cannot know that scoot and pick-and-place have similar effects, so it must learn this through embodied interaction with the world. We can apply research on learning from exploration and intrinsically motivated reinforcement learning (curiosity, exploration, and play) to enable an agent to explore a problem space in search of operator variants [PAED17, HS17, CBS05, CBCH16].

4. Discovery Detection. For the agent to have learned the *enclosed* predicate, it must detect unexpected events, which in this case was an unexpected change in a fluent. Appropriate mechanisms for tracking uncertainty using probabilistic approaches (e.g., Bayesian and information-theoretic methods), in combination with salience and attention mechanisms, can help detect these events. From the real-world standpoint, the agent must possess capabilities to be able to execute this exploration and discovery process, including grasping and manipulating unfamiliar

objects. These practical abilities are not trivial, and, in combination with intelligent reasoning, will provide a clear demonstration of agent autonomy while solving real-world problems.

5. Domain Extension. Finally, the agent must know how and when to assimilate new knowledge about its domain. This involves more than simply adding new domain elements; it must determine if the new knowledge will be consistent with existing content. Here, we can turn to research in belief revision that offers a formal approach to addressing these sorts of knowledge representation issues [HdMdBW14].

In addition, the agent must possess a set of crucial supporting skills, such as abilities to (1) dynamically invoke and use families of heuristics, (2) consider internal and external cues at varying levels of abstraction, (3) operate in a 'problem-stewing' or cue-monitoring mode, (4) formulate and redefine symbolic knowledge, including propositions, operators and goals, (5) perform meta-level searches over a problem space, and (6) learn and acquire knowledge from different domains. We believe that, to design agents that can solve MacGyver problems, it is essential that independent lines of research be merged. Importantly, we must combine the subtasks mentioned above and the supporting skills into an integrated cognitive system that can be embodied and situated in physical or virtual environments.

10.6 Evaluating Agents and Measuring Research Progress

Any proposed framework for intelligent agents must be accompanied by a discussion of how to evaluate those agents. For the Turing Test and many of its variants, one can only measure agents by the ultimate but subjective human judgment. This involves the question of being able to identify the source of behavior as human or artificial. Essentially, this was an all or nothing proposition that was highly subjective. We argue that this is insufficient. The MacGyver formulation offers subjective and objective options that were not available in many previous approaches. We propose three subclasses of measures: problem-centric , solution-centric, and agent-centric.

Problem-centric measures are based on a MacGyver problem's inherent diffi-

culty. Because such tasks have a domain-independent formulation, we can consider measures like the reachability of goal states and distance from start state, the size of the initial domain, the size of the minimal domain that contains both the start and goal state, and the existence and number of dead-end states (from which the agent can never solve the problem). Problem-centric measures, at their core, focus on the difficulty of MacGyver problems themselves, independent of the specific solution taken by the agent.

Solution-centric measures consider the solution found on a given MacGyver problem. Humans place high value on elegant and clever solutions to complex problems. We might quantify elegance and cleverness as complexity of an insightful strategy, the nature and number of domain transformations needed, the nature and number of heuristics and cues used in solving the problem, the time taken to solve a problem, the length of the solution plan, and in many other ways. Solution-centric metrics, when combined with problem-centric ones could even capture partial progress. For example, an agent making even limited progress on a very difficult problem might be scored higher than an agent that fully solves easier problems. Solution-centric measures could also capture subjective judgments of a human arbiter, such as judging whether a human or an artificial agent generated a solution.

Neither problem-centric nor solution-centric measures capture the inherent resource limitations of the agent and its environment. The agent's sensory-motor and cognitive capabilities limit how it initially represents problems, and which problems and strategies are viable. We would not expect a Roomba to pick up a block in cup world because it has no arms. Agent-centric measures might consider these resource limitations and therefore acknowledge that a task may be a MacGyver problem for one agent but not for another. Agent-centric measures could let us track and update the knowledge an agent has acquired during its lifetime. We can ask if the agent has improved its overall problem-solving capabilities and used knowledge gained from solving one task in another one. Finally, agents could be placed in adversarial games with one agent being tasked to design MacGyver problems for the other. This would let us evaluate an agent's creative ability not only to solve MacGyver problems, but

also to generate them.

We do not advocate for any one specific measure. Indeed, it might be best to take a hybrid approach that incorporates all three. The different metrics could let us study whether agents solving one MacGyver problem can solve similarly difficult (or easier) ones. We could also see whether the types of knowledge and heuristics that work with one class of MacGyver problems also work with another class.

Thus far, we have discussed evaluating problem-solving agents. However, we would also like to be able to track how the overall research community designs creative AI. To this end, we can track the progress on the independent subtasks identified earlier, which map well onto existing research agendas in AI subfields. In addition, we urge the community to invest time and funds to develop support skills needed for effective problem solving, such as the ability to draw on heuristics, to manipulate symbolic knowledge, and to play a role in integrated systems.

10.7 Conclusion

In this essay, we introduced a new class of AI challenges - *MacGyver problems* - defined as planning problems that are seemingly unsolvable, but where an agent can widen its representation of the domain, and maybe its understanding of the world, to discover solutions. Inspired by work in human insight problem solving and using the language of classical planning, we defined this class of tasks in a domain-independent framework that can generate domain-specific problem instances. We described one such instantiation using a toy domain called "cup world." We showed that recognizing and solving MacGyver problems are computationally intractable. Thus, a cognitive system must leverage families of heuristics, strategies, cues, and experiments to find solutions.

We walked through a heuristic approach to solving cup world, which also let us discuss the various agent capabilities needed to solve these problems. It is especially interesting to note the parallels between MacGyver problems and human insight problems, and we suggest that much can be learned from how humans solve

them. We also provided a short and nonexhaustive list of research subtasks, that is crucial to building cognitive systems that can solve MacGyver problems. Finally, we discussed possible ways to measure and evaluate agents that attempt to solve such problems, as well as research progress as a whole. The formulation does not require specific internal mechanisms or particular solution strategies, but instead focuses on general features of creative problem solving.

We believe the formulation of such challenge problems is only the first step, albeit a fruitful one, towards designing creative cognitive systems. We hope the formalism will help guide research by providing a set of formal specifications for measuring AI progress based on the ability to solve increasingly difficult MacGyver problems. We believe the cognitive systems community is ideally suited to tackle these challenge problems, and we encourage the community to use the framework and pursue research on (a) designing additional domain-specific instances of MacGyver problems, (b) updating cognitive architectures to solve these problems, and (c) conducting studies that compare performance of artificial agents to humans on these problems; even ones that are simple for humans like cup world might shed light on humans' heuristics.

Chapter 11

Solving MacGyver Problems

In Chapter 10, I introduced the idea of MacGyver problems and began the difficult task of formalizing creative problem solving as MacGyver problems in the language of classical planning. The formalisms presented in Chapter 10 were broad and covered problems encountered by one or more agents and covered domain transformations involving expansions and contractions. In this chapter, I conclude Part III with a discussion of algorithms that can be used to solve certain MacGyver problems. In doing so, I redefined certain aspects of the formalisms presented in Chapter 10 to make them simpler and more narrow in their scope.

11.1 Introduction

Let's begin by turning our focus to classical planning, the AI tradition that serves as a foundation for MacGyver problems. Classical planning involves building a sequence of actions, i.e., a plan, that achieves certain goals within an environment. Central to classical planning is the *domain-model*, a symbolic abstraction of the environment that captures objects, relationships and actions considered relevant to the goals. As we have seen from the last chapter, the domain model limits what goals a classical planning agent can conceivably achieve. Domain models are often handcrafted by domain experts and are intended to be general, allowing an agent to achieve a wide variety of goals within the environment. However, domain models can

be rendered unsuitable (goals unreachable) if the goals change or if the underlying environment changes in a way that requires the agent to consider *new* objects, actions or relationships. How can an agent adapt and transform its domain model when goals are unreachable, but discoverable? If future AI systems are to be resilient and autonomous in novel environments they must be able to continually do so during planning and acting.

Current approaches to addressing this problem, at the intersection of planning and learning, propose learning domain models [AFP+18] from observations [MZPS12], from action traces [GLP17], from reward signals [SAFMJ16], and from natural language [LRF+17]. A challenge that still remains is how to update domain models in real-time during problem-solving (planning and execution). In such cases, the agent will not only need to acquire new domain knowledge on-the-fly, unsupervised, but also apply it to the problem at hand immediately, and continue doing so until a goal is reached.

We saw that MacGyver problems are essentially classical planning problems, and are initially unsolvable for the agent because of the agent's incorrect or incomplete representation of its environment, either at the symbolic or the subsymbolic level. The MacGyver problem framework provides a suitable foundation to address the challenge of learning while problem solving, noted earlier. What makes solving MacGyver Problems difficult is the lack of reward signals, distance measures and heuristics, supervision, and a general lack of data to learn from. In this chapter, we propose a first approach for solving (or at least valiantly attempting) MacGyver problems.

We propose a set of algorithms (Section 11.3) for allowing an agent to fix an incorrect or incomplete classical planning domain by identifying anomalies (Section 11.3.2), adding new predicates (Section 11.3.3), generating new actions (Section 11.3.4), and generating new constant symbols (Sec: 11.3.5), in the absence of supervision, distance measures, and reward signals. The algorithms computationalize the ideas of sense-making and discovery inherent in human problem-solving. We demonstrate a proof-of-concept robotic implementation in a 3D physics-based environment.

11.2 The MacGyver Problem

Consider a first-order language $\mathcal{L}$ with atoms $p(t_1, \ldots, t_n)$ and their negations $\neg p(t_1, \ldots, t_n)$ (collectively, literals l), where t_i represents terms that can be variables or constants. A literal is grounded if and only if all of its terms are constants. We simplify and improve the MacGyver formalism introduced by Sarathy et al. [SS18b], as follows.

11.2.1 Defining a MacGyver Problem

Definition *(World).* World $\mathbb{W} = (S, A, \gamma)$ is a classical planning domain in $\mathcal{L}$. $S \subseteq 2^{\{\text{all ground atoms of } \mathcal{L}\}}$ denotes the set of states. A is the set of all ground instances a of operators in O, where O is the set of operators. An operator is a tuple $o = (n(x_1, \ldots, x_m), \text{pre}(o), \text{post}(o))$, where n is the operator symbol and x_i are variable symbols, and pre(o) and post(o) are sets of literals. $\gamma(s, a) = (s - \text{post}^-(a)) \cup \text{post}^+(a)$[1], if $a \in A$ is applicable to $s \in S$, and otherwise $\gamma(s, a)$ is undefined. S is closed under γ.

The World $\mathbb{W}$ describes the environment in which the agent is planning. An agent subdomain, defined over $\mathcal{L}_t$ is a part of $\mathbb{W}$ (or the entirety of the world itself) that is within the agent's awareness.

Definition *(Agent Subdomain).* An agent subdomain $\Sigma_t = (S_t, A_t, \gamma_t)$ is also a classical planning domain in $\mathcal{L}_t \subseteq \mathcal{L}$ that corresponds to the agent's perception and action within its world $\mathbb{W}$.

$$\Sigma_t = \{(S_t, A_t, \gamma_t) \mid ((S_t \subseteq S) \vee (A_t \subseteq A) \vee (\gamma_t \subseteq \gamma))\}$$

A MacGyver Problem is a classical planning problem that is defined in the World $\mathbb{W}$, but outside the agent's initial subdomain Σ_0. Formally,

Definition *(MacGyver Problem).* MacGyver Problem $\mathcal{P}_M = (\mathbb{W}, \Sigma_0, s_0, g)$, where Σ_0 is the initial agent subdomain and

[1] The + and − superscripts indicate subsets of post(o) corresponding the sets of positive and negative literals, respectively.

- $(S_0 \subset S) \vee (A_0 \subset A) \vee (\gamma_0 \subset \gamma)$
- $s_0 \in S_0$ is the initial state of the agent
- g is a set of ground atoms

Where $\nexists s \in S$, such that $s \in \hat{\Gamma}_{\Sigma_0}(s_0)$ and $g \subseteq s$

11.2.2 Solving MacGyver Problems

Definition *(Agent Domain Modification).* A domain modification Σ_u^* involves either a domain extension or contraction (we only consider extensions here). A domain extension Σ_u^+ of an agent is an Agent-subdomain at time u that is in the world $\mathbb{W}$ but not in the agent's subdomain Σ_t in the previous time t, such that $\Sigma_t \preceq \Sigma_u$. The agent extends its subdomain through sensing and perceiving its environment and its own self. E.g., the agent can extend its domain by making an observation, receiving advice or an instruction or performing introspection. Formally,

$$\Sigma_u^+ \;=\; \{(S_u^+, A_u^+, \gamma_u^+) \;\mid\; (S_u^+ \;\subseteq\; S \setminus S_t) \vee (A_u^+ \;\subseteq\; A \setminus A_t) \vee (\gamma_u^+ \;\subseteq\; \gamma \setminus \gamma_t)\}.$$

The agent subdomain that results from a domain extension is $\Sigma_u = \Sigma_t \cup \Sigma_u^+$. A domain modification set $\Delta_{\Sigma_t} = \{\Sigma_1^*, \Sigma_2^*, \ldots, \Sigma_n^*\}$ is a set of n domain modifications on subdomain Σ_t. Let Σ_Δ be the subdomain resulting from applying Δ_{Σ_t} on Σ_t.

Definition *(Strategy).* A *strategy* is a tuple $\omega = (\pi, \Delta_{\Sigma_t})$ of a plan π and a set Δ_{Σ_t} of domain modifications. A domain-modifying strategy ω^C involves at least one domain modification, i.e., $\Delta_{\Sigma_t} \neq \emptyset$.

Definition *(Context).* A *context* is a tuple $\mathbb{C}_i = (\Sigma_t, s_t)$ that represents the agent's subdomain and state at time t.

We are now ready to define an *insightful strategy* as a set of actions and domain modifications that the agent must perform for the problem's goal to be reachable.

Definition *(Insightful Strategy).* Let $\mathbb{C}_0 = (\Sigma_0, s_0)$ be the agent's starting context. Let $\mathcal{P}_M = (\mathbb{W}, \Sigma_0, s_0, g)$ be a MacGyver Problem for the agent in this context. An insightful strategy is a domain-modifying strategy $\omega^I = (\pi_t^I, \Delta_{\Sigma_t}^I)$ that, when applied in $\mathbb{C}_0$, results in a context $\mathbb{C}_u = (\Sigma_u, s_u)$, where $\Sigma_u = \Sigma_\Delta$ such that $\exists s' \in \hat{\Gamma}_{\Sigma_u}(s_u), g \subseteq s'$.

Intuitively, an insightful strategy captures idea that solving a problem involves exploration and domain expansion until the point (akin to the "Aha!" moment of insight) when finding a plan with the domain becomes feasible.

11.3 Proposed Algorithms

Our approach to solving MacGyver Problems is through hypothesis generation and active experimentation. We first define an oracle function so that we can formalize the process of requesting and using information from $\mathbb{W}$. While, the environment itself can be thought of as an oracle, the oracle could just as easily be any function that has access to state transitions in $\mathbb{W}$.

Definition *(Oracle).* Oracle $\varphi^{\mathbb{W}} : s_0 \times A \mapsto S$ has access to the transitions in $\mathbb{W}$ such that given a state $s \in S$ and an action $a \in A$, the oracle responds with a new state $s' \in S$ that need not be in S_t. We impose that the state presented to the oracle must be the state which an agent is currently in, namely its start state s_0.

Definition *(Planner)* A classical planner $\rho : \mathcal{P} \mapsto \Pi$ is a function that maps a classical planning problem $\mathcal{P}$ into a plan $\pi \in \Pi$ (namely, a sequence of actions) that is a successful plan or, if there is none, then returns $\emptyset$.

11.3.1 Overall Problem Escalation Procedure

Algorithm 11.1 directs the overall process by recursively decomposing the goal and escalating problem-solving. First, it attempts to solve using planner ρ, then escalates to explore other applicable actions with Algorithm 11.2, which then escalates further to create new actions (previously thought to be inapplicable) with Algorithm 11.4.

Algorithm 11.1 $solve(\mathcal{P}, \Delta_{\Sigma_t}, \varphi^{\mathbb{W}}, \rho, \Theta)$

1: $\mathcal{P} = (\Sigma_0, s_0, g)$: Planning Problem
2: Δ_{Σ_t}: The set of domain modification
3: $\varphi^{\mathbb{W}}$: oracle function
4: ρ: planner function
5: Θ: Exploration parameters
6: $\pi \leftarrow \rho(\mathcal{P})$
7: **if** $\pi \neq \emptyset$ **then**
8: **return** $\omega^I = (\pi, \Delta_{\Sigma_t})$
9: **else if** $\pi = \emptyset, |g| = 1$ **then**
10: $\Sigma_u^* \leftarrow hypothesize(\Sigma_t, s_0, \varphi^{\mathbb{W}}, \Theta)$, where $t < u$
11: $\Delta_{\Sigma_u} \leftarrow \Delta_{\Sigma_t} \cup \Sigma_u^*$
12: $\Sigma_u \leftarrow update(\Sigma_u^*, \Sigma_t)$
13: **return** $solve(\mathcal{P}_j, \Delta_{\Sigma_u}, \varphi^{\mathbb{W}}, \rho)$
14: **else if** $\pi = \emptyset, |g| > 1$ **then**
15: $g' \leftarrow g.remove(l \in g)$
16: **return** $solve(\mathcal{P}'(\Sigma_0, s_0, g'), \Delta_{\Sigma_t}, \varphi^{\mathbb{W}}, \rho) \cdot (\pi_{g' \rightarrow g}, \Delta_{\Sigma^t_{g'}})$
17: **end if**

Central to the approach is the process of sense-making by detecting anomalies, which are discrepancies between the current state and expectations imposed by the current domain model. The agent attempts to resolve existing anomalies by hypothesizing domain modification with Algorithms 11.2 and 11.3, or raise new anomalies (that it also subsequently resolves) by creating new actions.

Specifically, if ρ did not find a plan, Algorithm 11.1 recursively relaxes the goal. Line 16 shows the plan for achieving the original goal g from the state resulting from the execution of the plan for g', returned by the first recursive call. Once relaxed, it calls Algorithm 11.2 to hypothesize new actions.[2]

Running Example: Consider a running example in which an agent is given a domain definition for the classical 4-Op Blocks World domain extended with a predicate $visible(X)$. Now, suppose we deploy this agent in an environment with a cup and block (*Cups World*), and suppose the agent does not know any of the cup's affordances and believes it can apply the available actions to the cup, as well. Consider a planning problem $\mathcal{P}$ for this agent, with a goal $g = \neg visible(block)$, $\Sigma_0 =$ 4-Op Blocks World, and a start state, as follows:

[2]The agent does not know if $\mathcal{P}$ is a MacGyver Problem, which is an intractable problem in and of itself [SS18b].

$$s_0 = \{ontable(block), clear(block), ontable(cup),$$
$$clear(cup), handempty, visible(cup), visible(block)\}$$

In the interest of brevity, we will walk through only a single search for the solution. In reality, the agent might attempt several suboptimal choices before reaching the one we outline. Upon failure of ρ to solve the problem, it will escalate to Algorithm 11.2 for anomaly detection and domain modification.

11.3.2 Recognizing Anomalies

Algorithm 11.2 (*Hypothesize*) initializes sets P_1, P_{2A} and P_{2B} (initialize to $\emptyset$), which capture literals that are anomalies (described below). The algorithm also initializes a set *Selected* for tagging actions previously explored. While anomalies exist (line 8) (i.e., at least one of the sets P_1, P_{2A} and P_{2B} is not empty), it identifies applicable grounded actions in state s_0 that have not been previously selected (lines 9). If no applicable actions are available or if all actions have been tried, the algorithm escalates to Algorithm 11.4 to search for actions incorrectly thought to be inapplicable. If there are available applicable actions, the algorithm constrains the set into one that is smaller. The constraints can be based on meta-heuristics: for example, retaining only those actions that require an object to be grasped. Such a meta-heuristic might prevent an agent from trying irrelevant actions initially. In our running example, the agent will identify two applicable actions in s_0: $pickup(cup)$, $pickup(block)$.

The Algorithm 11.2 then randomly selects an action from the constrained set (line 14) and calls the oracle $\varphi^{\mathbb{W}}$ (i.e., executes the action, line 15), which brings it to the next state s_j. In our running example, suppose the agent chooses and executes $pickup(cup)$. Once, the action is performed, state s_1 follows:

$$s_1 = \{ontable(block), clear(block), holding(cup), visible(cup), visible(block)\}$$

Now, three types of anomalies can occur:

- A **Type I** anomaly occurs when the agent expects to observe a positive literal, but does not.

Algorithm 11.2 *hypothesize*$(\Sigma_t, s_0, \varphi^{\mathbb{W}}, \Theta)$

1: Σ_t: Domain
2: s_0: Start state (agent's current state)
3: $\varphi^{\mathbb{W}}$: oracle function
4: ρ: planner function
5: Θ: Exploration parameters
6: $P_1, P_{2A}, P_{2B} \leftarrow \emptyset$
7: $Selected \leftarrow \emptyset$
8: **while** $P_1 \cup P_{2A} \cup P_{2B} = \emptyset$ **do**
9: $app(a) \leftarrow \{a \in A_i^t \mid pre(a) \subseteq s_0, a \text{ is grounded}, a \notin Selected\}$
10: $A' \leftarrow constrain(app(a), \Theta)$
11: **if** $app(a) = \emptyset$ or $\forall a \in A', a \in Selected$ **then**
12: *break*
13: **end if**
14: Select $a \in A', Selected.add(a)$
15: $s_j \leftarrow \varphi^{\mathbb{W}}(s_0, a)$
16: **for all** $l \in post(a) \cup s_0 \cup s_j, l$ is a positive literal **do**
17: **if** $\exists l \in post(a) \setminus s_j$ **then**
18: $P_1.add(l)$
19: **else if** $\exists l \in (s_j \setminus s_0) \cap (s_j \setminus post(a))$ **then**
20: $P_{2A}.add(l)$
21: **else if** $\exists l \in (s_0 \setminus s_j) \cap (s_0 \setminus post(a))$ **then**
22: $P_{2B}.add(l)$
23: **end if**
24: **end for**
25: **if** $P_1 \cup P_{2A} \cup P_{2B} \neq \emptyset$ **then**
26: **return** $modify(a, P_1, P_{2A}, P_{2B}, \Sigma_t, s_j, \varphi^{\mathbb{W}}, \Theta)$
27: **end if**
28: $s_0 \leftarrow s_j$
29: **end while**
30: **return** $create(\Sigma_t, s_0, \varphi^{\mathbb{W}}, \Theta)$

- A **Type IIA** anomaly occurs when the agent observes a positive literal not in the previous state but is in the current state, not believed to be due to the current action.
- A **Type IIB** anomaly occurs when the agent observes a positive literal that was in the previous state is not present in the current state, not believed to be due to the current action.

Algorithm 11.2 performs a classification of anomalies and assigns the literals into three sets, each corresponding to an anomaly - P_1, P_{2A} and P_{2B}. Returning to our running example, the postconditions for the *pickup* action is $post(pickup(cup)) = \{holding(cup)\}$. For each of the literals in the $post(pickup(cup)$, we can determine the anomaly sets (lines 16-23), as $P_1 = \emptyset$,$P_{2A} = \emptyset$,$P_{2B} = \emptyset$.

Algorithm 11.2 then iterates to the next action. The two applicable actions are now $putdown(cup)$ and $stack(cup, block)$. Let's say the agent selected $stack(cup, block)$. Now, the postconditions are:

$$post(stack(cup, block)) = \{on(cup, block), clear(cup), handempty\}$$

However, an unexpected event occurs, which is the cup actually encloses the block. The agent does not have a concept or sense for this notion and will thus have to discover it.

$$s_2 = \{ontable(cup), clear(cup), handempty, ?clear(block), ?ontable(block), visible(cup)\}$$

We use "?" to signify an indeterminate predicate.[3] For this *stack* action, in contrast to the *pickup* action, the agent is going to catch the following anomalies (when comparing states s_2 with s_1 against $post(stack(cup, block))$: $P_1 = \{on(cup, block)\}$, $P_{2A} = \{ontable(cup)\}$, and $P_{2B} = \{visible(block)\}$

11.3.3 Modifying Action Definitions

After iterating through all the constrained actions, if there are any anomalies, Algorithm 11.2 calls Algorithm 11.3 to modify these actions. If there are no anomalies, it calls Algorithm 11.4 to consider inapplicable actions.

Algorithm 11.3 works through possible combinations of positive literals in the various anomaly sets. For anomalies of Type I, it removes the literal from the postconditions (line 10). For anomalies of Type IIA, it adds the literal to the precondition and adds a negation to the precondition (lines 11-12). For anomalies of Type IIB, it adds the literal to the precondition and its negation to the postconditions (lines 13-14). In our running example, the agent constructs a new action $a^*(cup, block)$ (variant of *stack*) and defines it as follows:

$$pre(a^*) = \{holding(cup), clear(block), visible(block), \neg ontable(cup)\}$$

$$post(a^*) = \{ontable(cup), \neg holding(cup), \neg clear(block), \neg visible(block)\}$$

[3] The agent, without additional knowledge, cannot know for sure if the block is still on the table as it cannot perceive it.

Algorithm 11.3 *modify*$(a, P_1, P_{2A}, P_{2B}, \Sigma_t, s_0, \varphi^{\mathbb{W}}, \Theta)$

Σ_t: Domain
s_0: Start state (agent's current state)
$\varphi^{\mathbb{W}}$: oracle function
ρ: planner function
Θ: Set of exploration parameters
for all $s_1 \in 2^{P_1}$ **do**
  for all $s_2 \in 2^{P_{2A}}$ **do**
    for all $s_3 \in 2^{P_{2B}}$ **do**
      $a' \leftarrow a$
      $pre(a').add(l, \forall l \in s_3)$
      $post(a').add(l' = \neg l, \forall l \in s_3)$
      $pre(a').add(l' = \neg l, \forall l \in s_2)$
      $post(a').add(l, \forall l \in s_2)$
      $post(a').remove(s_1)$
      if $test(a', \varphi^{\mathbb{W}})$ **then**
        return a' as a domain modification Σ_u^*
      end if
    end for
  end for
end for

Algorithm 11.4 *create*$(\Sigma_t, s_0, \varphi^{\mathbb{W}}, \Theta)$

Σ_t: Domain
s_k: Start state (agent's current state)
$\varphi^{\mathbb{W}}$: oracle function
ρ: planner function
Θ: Set of exploration parameters
$notapp(A) \leftarrow \{a \in A_i^t \mid pre(a) \not\subseteq s_0\}$
$A' \leftarrow constrain(notapp(A), \Theta)$
for all $a \in A'$ **do**
  for all $l \in pre(a)$ **do**
    $a' \leftarrow a$
    $pre(a').remove(l)$
    if $test(a', \varphi^{\mathbb{W}})$ **then**
      return a' as a domain modification Σ_u^*
    end if
  end for
end for

Algorithm 11.3 then calls $test(a', \varphi^{\mathbb{W}})$ to evaluate the proposed modification to the grounded action and attempt to generalize it. Informally, the $test(a', \varphi^{\mathbb{W}})$ function is a condition checker, which performs a series of experiments to evaluate the extent to which the modified action a' satisfies the several properties: repeatability (can a' be repeated from the same state?), token-invariance (does a' work with different object tokens of the same type, e.g., different cups?), reversibility (can a''s pre and postconditions be reversed?).

The agent can also make the action more compact by abstracting conjunctions of anomalous literals (i.e., those from P_1, P_{2A} and P_{2B}) into single predicates, via a rule such as

$$\phi(cup, block) \leftarrow ontable(cup) \wedge \neg on(cup, block) \wedge \neg visible(block)$$

This derived definition ϕ of the concept of "enclosure" might not be an exclusive one. For example, a transparent cup might not render the block invisible, however, it will still be enclosed. Our definition of "enclosure" via $\phi(cup, block)$ is still underdetermined (as is any concept learned inductively). For example, it cannot distinguish "covered" from "contained" (the latter being possible even when the cup is upright). Future work will explore how these concepts, discovered separately, can be unified using techniques in generalized planning [LSAJJ16].

11.3.4 Discovering New Actions

Algorithm 11.4 is called (if no anomaly is found) to explore the space of action possibilities that do not apply in the current state. It randomly relaxes preconditions to make actions applicable. If, in fact, the newly created action can be executed, it will be returned as an action modification that Algorithm 11.1 can use to solve the problem $\mathcal{P}_M$.

Returning to our running example, consider the following variant with not only abstract operators like *stack*, *pickup*, *putdown* and *unstack* in 4-Op Blocks World, but also sub-actions like *grasp*, *lift*, *moveAcross*, *setDown* and *release*, and sub-relations like *grasped*. Locations are represented with predicate $at(Object, Location)$. Note, decomposing actions to sub-actions and more generally exploring the space of hierarchically specified actions is an open research questions with recent promising results [GCS19]. Thus, for example, the $moveAcross(Object, Loc1, Loc2)$ action is defined to have preconditions $pickedup(Object)$ and $at(Object, Loc1)$, and post-condition $at(Object, Loc2)$. Now, consider a problem in which the start state only contains one block (and no cup) and the goal is to move the block from location-1 to location-2 *without lifting it*.

Upon exploration with Algorithms 11.1 and 11.2, the agent does not discover

any anomalies. This is not surprising as there is only one block and a limited set of actions. Consequently, no modifications are apparent. Algorithm 11.4 executes existing inapplicable actions.[4] Suppose the agent selects the *moveAcross* action and removes the precondition *pickedup* to create a new action b^* (which can be thought of as a "slide" or "nudge" action). When tested, the agent will be able to discover that the action sequence *grasp*, *lift*, *moveAcross*, *setDown*, and *release* has the same set of postconditions as *grasp*, b^*, *release*, thereby solving the $\mathcal{P}_M$ defined earlier.

11.3.5 Generating New Constants

Consider a variant of Cups World (Fig. 11.1) in which the agent must move exactly one (blue) of two blocks co-located at location l_1. Consider an agent that knows the actions *cover* and *nudge* discovered in the previous sections. Initial execution of the plan will also result in illegally moving the red block.

11.3.5.1 Diagnose

As before, the agent diagnoses the problem by stepping through each of the actions and identifying anomalies. It notices that covering the blue block also covers the red block, a Type IIB anomaly. Upon correction, the original problem is turned into a symbolic MacGyver problem. No *modification* or action *creations* discussed earlier will help. Agent cannot solve this problem by learning from action traces [AFP+18, GLP17] as no action trace will provide the needed knowledge. The agent must explore the subsymbolic space.

11.3.5.2 Reason and Define New Goal

We propose that the agent reasons over its symbolic knowledge to determine key constraints that can limit the subsymbolic search: (1) which low-level (primitive) actions to explore, (2) which parameters of these actions to vary, and (3) what is the termi-

[4]Note, the agent might need some set of practical and normative constraints, to limit what actions it can and cannot create so as to prevent the agent from damaging itself or others.

nating condition or goal of the subsymbolic search. The earliest point in time when a Type IIB anomaly was spotted was after the performance of the *cover* action, when the red block disappeared unexpectedly. We consider actions at and before this time that might have led to the anomaly, namely, *cover*, *move-across*, *lift*, *grasp* as possible candidates and backtrack through them. The anomalous atom was $visible(?x)$ and so we consider those groundings of the atom among all possible groundings of this atom that also match the desired criteria for the *cover* when the anomaly occurred. Between the two blocks, there are eight possible groundings. However, we only consider the two groundings in which the blue block is rendered invisible as this is a requirement for us to be able to move the block without touching it: (1) $\neg visible(b), \neg visible(r)$ and (2) $\neg visible(b), visible(r)$. We know from the initial attempt that grounding #1 is not going to work. Thus, the agent adopts grounding #2 as a new goal.

11.3.5.3 Exploration and Domain Update

The agent searches (we employ a straightforward binary search) through the subsymbolic space of *cover* by translating the cup at varying vertical distances and observing whether the new goal is satisfied. If after some number of tries, the agent does not succeed, it backtracks to the *move-across* action and translates the cup along various horizontal positions. Crucially, for each position, the agent executes not only the *move-across* action but also all subsequent actions until and including *cover*. The agent discovers that translating the cup to approximately half way between locations l_1 and l_2 works. Our agent has thus far maintained a database of symbolic locations keyed to location names and valued to the x, y, z coordinate in space. The agent introduces a new location $l_{1.5}$ into this database with its position. Once introduced, this symbolic location is added into the symbolic domain and the agent can execute its previous plan modified with the new location. Now, the agent has not yet learned any properties associated with $l_{1.5}$. However, performing the sequence of actions *cover*, *nudge* and *uncover*, will provide the agent with an opportunity to learn that this new location is adjacent to both l_1 and l_2. Studying

such incidental learning opportunities is the subject of future work.

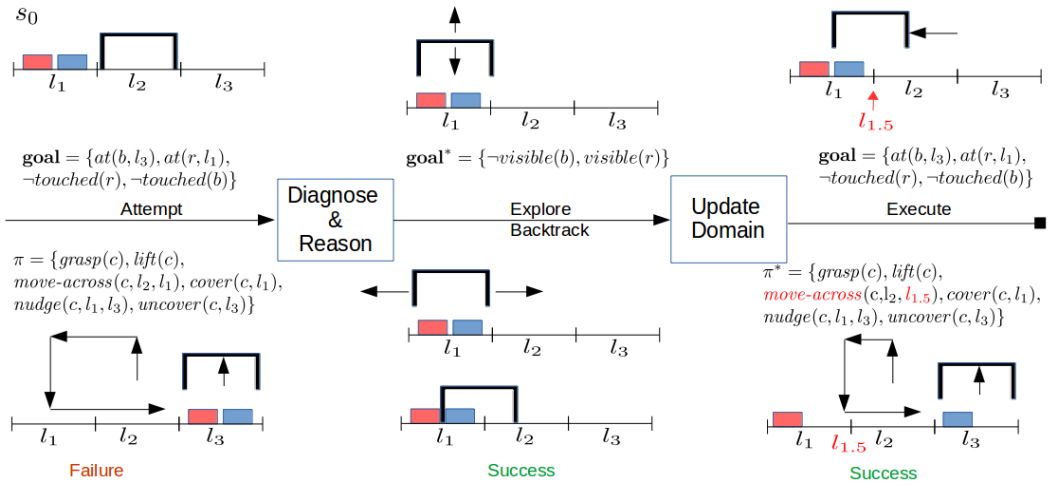

Figure 11.1: Problem solving when symbolic-macgyvering fails. The agent must explore an infinitely large subsymbolic space through exploration. We demonstrate an approach in which symbolic reasoning can provide constraints for a subsymbolic search over an action parameter space. The agent, capable of moving the blue block (b) without touching it (original Cups World), is now prohibited from touching or moving the red block (r). The agent learns, through reasoning, exploration, and noticing anomalies that moving the cup across to a new location between l_1 and l_2 will solve this variant of Cups World.

	Nudge		Enclosure	
Run	# **c**	**t**	# **c**	**t**
1	12	72	22	162
2	10	66	21	146
3	5	27	12	70
4	5	32	6	38
5	9	67	23	125
6	3	17	18	107
7	4	24	6	49
8	4	26	11	73
9	10	69	22	149
10	7	43	12	64
Avg	**6.9**	**44.3**	**15.3**	**98.3**

Table 11.1: Number of Oracle calls (c) and time (t) taken (in sec) to discovering "nudge" and "enclosure" in Blocks World and Cups World. Trials performed by a Fetch robot in simulation using algorithms integrated into the DIARC cognitive robotic architecture.

11.4 Robot Integration and Experiments

There is currently no agreed-upon protocol for evaluating MacGyver Problems. These are open-ended creative problems, for which evaluation is unsurprisingly quite challenging. Nevertheless, we show a first robotic implementation of a MacGyver agent and some initial experimental results in a 3D robotic simulation environment.

To integrate creative problem solving into a robotic system, we developed several architectural components for performing certain key operations – (1) hypothesis generation (goal and action decomposition, pre-/postcondition relaxation), (2) action execution, planning, and testing protocols, and (3) sense-making (sensor parsing, noticing anomalies). We built these components within two robotic frameworks, each capable of maintaining symbolic knowledge of a planning domain as well as mappings between the planning representations and low-level action and relational primitives.

We implemented these components in the Agent Development Environment (ADE), which is a robot-agnostic middleware for cognitive architectures [AS06]. We also implemented the approach in ROS, an open-source robotic framework, with ROSPlan [QCG+09, CFL+15], a framework for embedding a generic planner in a ROS system. We selected off-the-shelf classical planners for the planner ρ, and Gazebo (a ROS-based 3D physics simulation environment) for our oracle $\varphi^{\mathbb{W}}$. We implemented the $test(a', \varphi^{\mathbb{W}})$ within this framework in a limited way allowing agents to repeat actions.[5]

We selected the Fetch Robot[6] and attempted to characterize (at least roughly) the quality of performance on our running example (Cups World) over several runs in simulation both for discovering enclosure (Sections 11.3.2 and 11.3.3) as well as creating the nudge action 11.3.4. All simulations and experiments were run on a desktop with 12 cores, each an Intel® Core™ i7-5820K 3.30GHz processor and 31.3 GB of RAM. The results in Table 11.1 show that the robot was indeed able to learn the missing actions and solve MacGyver Problems in a reasonable amount of time.

[5] We will release our ROS-based code on Github.
[6] Robotic Platform by https://fetchrobotics.com/robotics-platforms/ with ROS bindings.

Since there are no general heuristics for solving these problems, we can only attempt to find and characterize heuristics that cover large classes of problems. Doing so will require a systemic exploration based on the framework and base architecture provided in this chapter.

11.5 Other Related Work

This work overlaps with many subfields in AI. First, it builds on recent efforts by Sarathy and Scheutz in formalizing MacGyver Problems [SS18b]. While describing a possible solution strategy using the Cups World domain, they did not present algorithms for solving MacGyver problems. A related field of contingent planning focuses on allowing an agent to plan online as it senses an environment [KS16a]. Some have proposed algorithms for planning with incomplete information, whereby conditional plans have to be developed [BG00, AFP+18]. However, in both these approaches, the set of possible states or belief states are assumed to be known to the agent, which is not the case in MacGyver Problems. Recent work in using deep neural networks have shown promise in learning planning domain models from images [AF18, APA+18]. However, even if the agent is able to learn the planning domain, a challenge still remains as to how it can revise these representations when planning fails, as is required for MacGyver Problems. Some work in belief revision [Her14] might be useful here, and we have ourselves employed notions of goal reformulation in our approach.

Integrated approaches for growing a symbolic representations with what the agent can learn from exploring the subsymbolic space shares similarities with our approach to learning new constants [KLP13, KKLP14, SFR+14, KKLP18]. While these approaches are not configured to solve MacGyver problems, they nonetheless provide promising techniques to enhance our subsymbolic search capabilities.

Reinforcement learning offers a helpful way to model exploration and learning. Two promising approaches relevant to creative problem solving are (1) curiosity-driven or intrinsically motivated RL, and (2) hierarchical RL. The first approach

attempts to handle domains were reward functions (even long range ones) are not readily available [CBS05, HS17]. There is a parallel here to MacGyver Problems which is the lack of reward. However, those RL methods still do not allow an agent to modify its own action parameters. Hierarchical RL and its use of "options" appears promising for representing meta-level strategies [SPS99, KB09, KKLP18]. Segovia-Aguas et al. [[SAFMJ16]] proposed a combined planning and RL approach for learning low-level action policies. However, this work assumes a complete high-level model, which we do not have for MacGyver problems.

The notion of MacGyvering has been explored by researchers in robotics and tool creation [NSZC19], demonstrating techniques for learning novel object affordances through interaction and exploration. While similar in spirit to the proposed approach, these approaches are quite different in that they are focused on solving specific geometric problems of tool construction in the robotic domain. In contrast, the proposed approach is aimed at a higher level of abstraction, generality and domain-independence.

11.6 Open Problems

We have presented an approach to creative problem solving that actively involves environmental exploration, similar to real-world human problem solving [Sar18b]. Indeed, by definition, MacGyver Problems require such exploration, without which the planning task is unsolvable. In addition, we computationalize ideas of constraint relaxation, goal decomposition, and anomaly detection from psychology and insight problem solving literature [LJ88, Ohl92]. As noted earlier, there are no domain general heuristics or even heuristic discovery algorithms available for MacGyver Problems. Thus, an agent solving such problems must rely on some form of organic exploration, observation and domain transformation in order to build a path from the start state to a goal.

The proposed approach will guarantee domain modifications if they are permitted in the environment in which the agent exists. One major challenge is the

possibility of entering dead ends and traps. While that is possible, the algorithm is inherently designed to create new actions and dynamically adapt its domain model, which in turn has the potential for it to discover a path out of a trap. We have yet to evaluate this capability, but plan to do so in future work.

These are one of many important steps needed for designing resilient and autonomous agents. There are several limitations to the proposed approach, a significant one being computational complexity. Iterating through various actions and possible arrangement of pre and postconditions is at best of exponential complexity, no better than classical planning itself. However, meta-heuristics on the allowed types of actions might help limit this search space - e.g, through the use of $constrain(A, \Theta)$ function mentioned in the algorithms. One example of a meta-heuristic might be: when attempting to solve a problem involving objects, it is best to explore actions that use and/or manipulate these objects. In general, this question of smart exploration could interface with the growing body of literature in exploration strategies for reinforcement learning. We propose a model for using symbolic reasoning to help constrain subsymbolic exploration. It is still an open question as to whether such an approach could help an agent make its exploration and learning more sample efficient. If the agent used a collection of heuristics to simplify the search space, there are many open questions as to the space and nature of these heuristics.

Although we have discussed MacGyver problems as single isolated problems, in reality, the agent might encounter them during the course of a curriculum [NPL+20]. In fact, the agent may encounter a training regimen specifically designed with increasingly difficult macgyver problems. In such cases, there are open questions about how the agent can improve its performance between problems. Can it learn general problem solving techniques applicable not just across different tasks in the same domain, but potentially different but overlapping domains? The agent might need to build meta-abstractions over its current domain abstractions to capture general properties of domain transformations while solving particular MacGyver problems. By scaffolding the agent's learning in a systematic, incremental curriculum, it may be possible to make this process of building meta-abstractions (i.e., those

capturing problem-solving heuristics) more tractable. An example of such a curriculum could involve teaching an agent about finding hidden objects. If the agent solves a problem involving a cup enclosing another object, it might be able to learn that cups can enclose things (i.e., a particular affordance of a particular class of objects). But, it might also learn that objects that are not visible could be hidden by other objects and can be discovered by moving objects around (i.e., a meta-level problem solving technique).

11.7 Conclusion

Future AI systems will need to able to adapt their own mental model of the environment dynamically as they explore an ever-changing world. They must be able to extend it with limited data and without any external reward signal or supervision. In this chapter, we propose the first steps for designing such creative problem-solving systems. Specifically, we provide an approach for solving a class of problems, called MacGyver Problems, which are seemingly impossible at first, but ultimately solvable through domain modifications. We show how symbolic planning can be improved through exploration and anomaly detection, and how subsymbolic planning can be guided by high-level reasoning. The potential for problem-solving through self-guided exploration and discovery is great. Not only can AI systems solve difficult problems in isolation, but they can work in teams to explore different parts of the problem-space and accelerate problem-solving.

Chapter 12

Conclusions

In the introduction, I began by posing the following question: *what would it take for an artificial intelligence (AI) agent to appreciate the humor in a classic Harry Bliss New Yorker cartoon?*. Throughout this dissertation, I have endeavored to argue that extracting statistical patterns from data, by itself is not sufficient for developing AI systems that can understand human dialogue, use and manipulate objects, abide by our human norms or even think creatively, let alone understand and leverage all the hidden assumptions in the cartoon. What is needed is the ability to represent (and *reason* with) assumptions about the physical, cognitive and social aspects of our world.

Making and breaking these kinds of **assumptions** is crucial to human intelligence and creativity. They fill gaps in our knowledge and resolve ambiguities in an otherwise endless stream of perceptual information. When broken, they allow us to restructure our knowledge in new ways and discover previously unknown connections. Handling assumptions is required not only for understanding humor, but making sense of most everyday experiences and interactions. I contend that future AI systems must also be able to ***reason with implicit assumptions*** if they are to work effectively with humans.

The problem, however, is that many of the necessary common sense assumptions are elusive and remain hidden in our collective minds. It is also unclear what kind of reasoning is needed to effectively use these assumptions while thinking.

I have sought to uncover the computational nature of these hidden assumptions and how we make and break them. My approach has been to (1) identify AI domains that can benefit from reasoning with hidden knowledge, (2) formalize the types of knowledge and reasoning needed within these domains, (3) develop efficient algorithms for extracting the hidden assumptions, (4) develop cognitive architectures for reasoning with these assumptions, and (5) design studies to validate the effectiveness of new methods in practical settings. Developing effective approaches requires bringing together many disciplines covering areas of knowledge representation and reasoning, machine learning, natural language processing, uncertainty processing, human-robot interaction, as well as cognitive science, social psychology and ethics. Below, I discuss my contributions and outline future directions of research.

12.1 Language Understanding

12.1.1 Using hidden assumptions for resolving pronouns

Consider the discourse: "Smith entered the office of his boss. He was nervous."[1] We might assign the "he" pronoun in the second sentence to Smith, assuming naturally that he is nervous because he is getting reprimanded, an interpretation that makes sense. However, if our next sentence reads: "After all, he did not want to lose his best employee," we immediately revise our prior conclusion. This is an example of **nonmonotonic reasoning** and is pervasive in human language understanding. To handle such sentences, I introduced a new class of problems, which requires tracking state changes as language unfolds and reasoning about possible worlds within these intermediate states [SS19b]. To resolve pronouns of this form, I proposed a multi-reasoner approach that employs Answer Set Programming, a declarative programming paradigm that permits nonmonotonic reasoning. I showed that current state of the art neural systems fail to correctly resolve even very seemingly simple discourses such as "Pick up block-A. Put it on block-B. Pick up block-C. Put it on block-B."

[1] This example is courtesy Grigris Antoniou.

12.1.2 Using hidden assumptions for understanding speech acts

More recently, I extended the same architecture to resolving **indirect speech acts** (ISAs). These are utterances whose meaning might appear somewhat inconsistent with the surface form. For example, the question "Can you describe the shark?" might be requesting a shark description, and not asking the listener for a yes or no answer. Thus, the utterance has an indirect meaning of a request rather than an ask or a question. These are common politeness norms in human communication, and AI dialog systems must also be able to handle them correctly. However, much like the pronoun resolution problem mentioned earlier, ISAs also require nonmonotonic reasoning. I developed such a model that includes many domain-independent default rules and complex nonmonotonic reasoning patterns. I am currently developing a more challenging corpora for evaluating the model and interpreting ISAs in different contexts.

12.1.3 Future work

Like pronouns and ISAs, there are many linguistic phenomena (e.g., word sense and presuppositions) that are fraught with ambiguity, and resolving which requires nonmonotonic reasoning with hidden knowledge. In future work, it will be fruitful to develop language datasets that can be used to jointly test and compare neural-based as well as logic-based approaches. These datasets can then be used to evaluate the design of architectures for simultaneously resolving several forms of linguistic ambiguities, a capability offered by my multi-reasoner approach. Morevoer, the reasoning and inference machinery that comes with logic-based approaches can be used to extract hidden knowledge from text (e.g., online instructional datasets like WikiHow) as well as neural language models (e.g., BERT). Many real-world domains, like legal reasoning, in which extracting the hidden knowledge is particularly challenging even for humans, could benefit from integrating logical-reasoning-based machinery. In this domain, norm-based and nonmonotonic reasoning systems can help us evaluate the strength of legal arguments by extracting hidden assumptions.

It is desirable to design a unified architecture capable of simultaneously resolving different ambiguities at different levels of language understanding (word, sentence, discourse etc.). "Understanding" then is a matter of choosing the interpretation that makes the most sense across a variety of hidden assumptions and background knowledge.

12.2 Common Sense Assumptions about Social Norms

One category of hidden common sense assumptions are those associated with human **social and moral norms**. Norms play an important role in society in maintaining order and ensuring cooperation and coordination. However, norms are challenging to computationalize as they are highly context-dependent, culture-dependent, and dynamic. Here, I will discuss my contributions in *modeling human norm acquisition and building AI agent architectures for norm learning and reasoning.*

12.2.1 Social Norms Associated with Objects and Interactive Behaviors

Norms for object affordances: There is a tight connection between an object's physical appearance and the triggering of action possibilities in our minds. This allows us to quickly choose and manipulate an object correctly and effectively. When we see a knife, we also see possibilities for how and where we can grasp it. While visual and geometric features of objects might serve as useful cues for inferring basic actions like grasp, push, and pull, they do not tell us the complete story. For example, an antique knife behind a glass case in a museum does not allow for the same sort of grasping action. Other contextual aspects combined with our common sense social norms of the etiquette in a museum allow us to infer correctly. I introduced the formal notion of "**cognitive affordances**"[2] to capture these types of more complex normatively-charged action possibilities [SS16c, SS15b]. I proposed a probabilistic logic-based formalism for reasoning about social norms associated with tools in a

[2] "Affordances" generally represent the relationship between an object and the set of actions an agent can perform on or with the object (e.g., a mug has a grasp affordance).

kitchen domain. I integrated this formalism into a cognitive robotic architecture that allowed for evaluating the performance on a PR2 robot [SS16d, SS16a]. The robot was able to consider **implicit social norms** associated with handling sharp objects. It reasoned that when using a knife, it should grasp the handle, whereas when handing the knife over to someone, it should grasp the blade, allowing for safe transfer.

Consent norms for robots: Certain normative assumptions extend beyond specific objects or domains and permeate all forms of interaction. One such example is the notion of "**implicit consent**," the idea that certain interactive behaviors demand prior implicit permissions. Consent (or lack thereof) can alter the normative valence of a situation, turning appropriate behaviors into inappropriate ones. I turned to centuries-old legal precedent on consent in tort law and suggested how these established legal principles can guide robotics researchers while designing ethical behaviors and conducting usability studies [SAS19].

12.2.2 Generalized Norm Representation and Learning

Learning norms from language: Where do these social norm assumptions come from? One source is natural language instruction. I designed algorithms that allow a robot to learn, from natural language instruction received from single or multiple speakers, the social rules associated with handling knives [SOKS18]. Not all norms are learned from language, however. Many are learned simply from observation. Observing "norms" is tricky, as what one is observing is often not a norm *per se*, but simply the performance (or non-performance) of an underlying behavior.

Modeling human norm learning: I explored the nature of human norms and their cognitive representations [Sar19]. Crucially, I showed that these mental representations actually encode contextual aspects and inherently capture epistemic uncertainty [SSA+17]. I performed several human-subject studies and developed a preliminary computational model for human norm learning under uncertainty. The model was built on a logical framework and coupled with Dempster-Shafer theory (DST), a generalization of the Bayesian approach to learning under uncertainty. I

extended this work to enable agents to represent their uncertainty over contexts as well [SSM17].

Learning norms at scale: A limitation of DST is its computational complexity. I developed a set of generalized learning algorithms for reducing this complexity. I applied these algorithms to large-scale agent-based simulations during which an agent learned several norms from observing hundreds of other agents. This work extended the current literature on multi-agent norm learning by showing how norms can be learned under more realistic conditions where the learning is influenced by incomplete, imperfect and unreliable data.

12.2.3 Future work

There are many opportunities for future work in better understanding how social norms can be incorporated into AI systems. One direction is to apply normative principles associated with consent in specific domains where consent can influence robot behaviors. For example, it would be useful to identify the visual and linguistic consent cues suggestive of object ownership norms. Knowing when it is appropriate to borrow objects and tools is important for robots engaged in collaborative teamwork. Robotic architectures capable of detecting these cues and adjusting their behaviors to conform with these norms could be a fruitful direction of future research.

Exploring how machines can learn in open-world settings under uncertainty is another potentially rich area of future research. I discussed the use of a Dempster-Shafer theory (DST) in connection with learning norms under uncertainty. An advantage of DST is that it has the mathematical ingredients for "**open-world learning**," in which the space of possible behaviors (over which probability distributions are learned) is unknown at the start. An agent equipped with this capacity will be able to assimilate and establish the normativity of new behaviors on the fly while learning. Open-world agent architectures and algorithms can be designed that apply these powerful mathematical ingredients in DST and allow agents to learn from a state of complete ignorance, not only of the normative status of behaviors, but of the knowledge of the behaviors themselves.

12.3 Breaking Assumptions: From Common Sense to Creativity

I have discussed the importance of common sense assumptions as gap fillers that allow us to make sense of situations. Sometimes, these common sense gap filling assumptions can misguide our perception of the world. Illusionists and magicians frequently use this aspect of human cognition when designing tricks. Cartoonists, like Harry Bliss, also do so with drawings and text that seem incongruent with one another, but when combined suggest a better comic model. Human cognition is flexible and can break existing assumptions and adopt new ones in the face of new knowledge.

12.3.1 Formalizing creative problem solving

This occurs frequently in the real world, whether it is a mundane task of discovering a new affordance for an object (e.g., using a chair to hold open a door) or whether it is the historic task of a brilliant scientist proposing a paradigm shifting theory. In all these cases, we are (and can be) made to rethink our common sense assumptions and reframe them as needed. Undoubtedly, this is an important skill for AI systems as well, especially if they are to assist us with some of humanity's toughest challenges requiring major paradigm shifts. I recently proposed a class of problems, dubbed "**MacGyver Problems**," which are designed to seem initially unsolvable for an agent, but ones that an agent can conceivably solve by interacting with its environment [SS18c]. I have also begun developing algorithms for solving certain limited classes of MacGyver problems (Chapter 11). MacGyver problems closely parallel insight problem solving, an area that has been studied by psychologists for over a century. I proposed a new cognitive model for creative problem solving that extends prior work on insight problem solving and is supported by neural evidence [Sar18b].

12.3.2 Future work

Extend the theoretical framework for MacGyver problems by defining complexity classes and designing evaluation frameworks for testing new solution algorithms is an exciting avenue for research into creative AI systems. The theoretical framework offers us the ability to systematize the process of generating novel stimuli for studying the neural basis of insight problem solving as well, an area ripe for collaboration with neuroscientists.

12.4 In Conclusion

At the core of the process of making and breaking assumptions is the ability to continually assess and make sense of the world, resolving ambiguity to favor explanations that make the most sense, and recognizing anomalies as possible avenues for creative exploration. Agents designed with the ability for sense-making are likely to be more robust, resilient, resourceful, trustworthy and creative.

Bibliography

[18991] O'brien v. cunard steam ship co. In *28 N.E. 266*, 1891.

[19880] Commonwealth v. appleby. In *402 N.E.2d 1051*. Mass: Supreme Judicial Court, 1980.

[20112] Commonwealth v. carey. In *463 Mass. 378*. Mass: Supreme Judicial Court, 2012.

[ABB+04] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, (4):1036, 2004.

[ABMMT15] David Abel, Gabriel Barth-Maron, James MacGlashan, and Stefanie Tellex. Affordance-Aware Planning. 2015.

[Abr13] Anna Abraham. The promises and perils of the neuroscience of creativity. 7(June):1–9, 2013.

[ACC+07] Giulia Andrighetto, Marco Campennì, Rosaria Conte, Mario Paolucci, Labss Istituto, and S Martino. On the Immergence of Norms : a Normative Agent Architecture The Intra-agent Processes : EMIL-A. *Emergent Agents and Socialities: Social and Organizational Aspects of Intelligence. Papers from the 2007 AAAI Fall Symposium*, pages 11–18, 2007.

[ACCC10] Giulia Andrighetto, Marco Campenni, Federico Cecconi, and Rosaria Conte. The Complex Loop of Norm Emergence: A Simulation Model. pages 19–35, 2010.

[Ack16] Jennifer Ackerman. *The genius of birds*. Penguin, New York, NY, 2016.

[AF14] John R. Anderson and Jon M. Fincham. Discovering the sequential structure of thought. *Cognitive Science*, 38(2):322–352, 2014.

[AF18] Masataro Asai and Alex Fukunaga. Classical planning in deep latent space: Bridging the subsymbolic-symbolic boundary. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[AFP+18] Ankuj Arora, Humbert Fiorino, Damien Pellier, Marc Métivier, and Sylvie Pesty. A review of learning planning action models. *The Knowledge Engineering Review*, 33, 2018.

[AH81] Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[AKS17] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. Value alignment or misalignment – what will keep systems accountable? In *AAAI Workshop on AI, Ethics, and Society*, 2017.

[AL01] Nicholas Asher and Alex Lascarides. Indirect speech acts. *Synthese*, 128(1-2):183–228, 2001.

[AMC14] Jacopo Aleotti, Vincenzo Micelli, and Stefano Caselli. An Affordance Sensitive System for Robot to Human Object Handover. *International Journal of Social Robotics*, 6(4):653–666, 2014.

[AML16] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *Proceedings of the Workshop on AI, Ethics, and Society at the the 30th AAAI Conference on Artificial Intelligence*, 2016.

[APA+18] Leonardo Amado, Ramon Fraga Pereira, Joao Aires, Mauricio Magnaguagno, Roger Granada, and Felipe Meneguzzi. Goal recognition in latent space. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018.

[AS06] Virgil Andronache and Matthias Scheutz. ADE: An architecture development environment for virtual and robotic agents. *International Journal of Artificial Intelligence Tools*, 15(2), 2006.

[AS17a] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[AS17b] Pascal Amsili and Olga Seminck. A Google-proof collection of French Winograd Schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, 2017.

[AS17c] Thomas Arnold and Matthias Scheutz. Beyond moral dilemmas: Exploring the ethical landscape in hri. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 445–452. IEEE, 2017.

[AS18a] Thomas Arnold and Matthias Scheutz. Observing robot touch in context: How does touch and attitude affect perception of a robot's social qualities? In *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[AS18b] Thomas Arnold and Matthias Scheutz. Observing robot touch in context: How does touch and attitude affect perceptions of a robot's social qualities? In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 352–360. ACM, 2018.

[AS19] Ekaterina Abramova and Marc Slors. Mechanistic explanation for enactive sociality. *Phenomenology and the Cognitive Sciences*, 18(2):401–424, 2019.

[Asa07] Peter M Asaro. Robots and responsibility from a legal perspective. *Proceedings of the IEEE*, pages 20–24, 2007.

[ASF14] John R Anderson, Hee Seung, and Jon M Fincham. NeuroImage Discovering the structure of mathematical problem solving. *NeuroImage*, 97:163–177, 2014.

[Aus75] John Langshaw Austin. *How to do things with words*. Oxford university press, 1975.

[AW06] Ivan K Ash and Jennifer Wiley. The nature of restructuring in insight: An individual-differences approach. *Psychonomic Bulletin & Review*, 13(1):66–73, 2006.

[BA03] Ergun Bicici and Robert St Amant. Reasoning About the Functionality of Tools and Physical Artifacts. *Department of Computer Science, North Carolina State University, Tech. Rep*, 22:1–34, 2003.

[BAB06] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2006.

[Bar03a] Chitta Baral. *Knowledge representation, reasoning and declarative problem solving*. Cambridge university press, 2003.

[Bar03b] Chitta Baral. *Knowledge representation, reasoning and declarative problem solving*. Cambridge university press, 2003.

[BAS15] Rahmatollah Beheshti, Awrad Mohammed Ali, and Gita Sukthankar. Cognitive Social Learners: An Architecture for Modeling Normative Behavior. *Twenty-Ninth AAAI Conference on Artificial Intelligence Cognitive*, pages 2017–2023, 2015.

[BB09] A K Barbey and L W Barsalou. Reasoning and Problem Solving : Models. 8:35–43, 2009.

[BBF01] Selmer Bringsjord, Paul Bello, and David Ferrucci. Creativity, the Turing test, and the (better) Lovelace test. *Minds and Machines*, 1:3–27, 2001.

[BBS15] Abdeslam Boularias, J Andrew Bagnell, and Anthony Stentz. Learning to Manipulate Unknown Objects in Clutter by Reinforcement. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Learning*, pages 1336–1342, 2015.

[BBSS16] Roger E Beaty, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative cognition and brain network dynamics. *Trends in cognitive sciences*, 20(2):87–95, 2016.

[BBW+14] Roger E Beaty, Mathias Benedek, Robin W Wilkins, Emanuel Jauk, Andreas Fink, Paul J Silvia, Donald A Hodges, Karl Koschutnig, and Aljoscha C Neubauer. Creativity and the default network: A functional connectivity analysis of the creative brain at rest. *Neuropsychologia*, 64:92–98, 2014.

[BC17] Siddhartha Banerjee and Sonia Chernova. Temporal models for robot classification of human interruptibility. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1350–1359. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[BdCPD+13] Tina Balke, Célia da Costa Pereira, Frank Dignum, Emiliano Lorini, Antonino Rotolo, Wamberto Vasconcelos, and Serena Villata. Norms in mas: Definitions and related concepts. In *Dagstuhl Follow-Ups*, volume 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[BDK16] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. In *Springer handbook of robotics*, pages 1935–1972. Springer, 2016.

[Bel10] Sieghard Beller. Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. *Cognitive Processing*, 11(2):123–132, 2010.

[Ben15] David Bender. Establishing a human baseline for the Winograd Schema Challenge. In *MAICS*, pages 39–45, 2015.

[BG00] Blai Bonet and Hector Geffner. Planning with incomplete information as heuristic search in belief space. In *Proceedings of the Fifth International Conference on Artificial Intelligence Planning Systems*, pages 52–61. AAAI Press, 2000.

[BHL+15] Dan Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. The Winograd Schema challenge and reasoning about correlation. In *Proceedings of the AAAI Spring Symposium on Symposium on Logical Formalizations of Commonsense Reasoning*, 2015.

[Bic06] Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, New York, NY, 2006.

[BJB+16] Mathias Benedek, Emanuel Jauk, Roger E Beaty, Andreas Fink, Karl Koschutnig, and Aljoscha C Neubauer. Brain mechanisms associated with internally directed attention and self-generated thought. *Scientific reports*, 6(October 2015):22959, 2016.

[BJF+14] Mathias Benedek, Emanuel Jauk, Andreas Fink, Karl Koschutnig, Gernot Reishofer, Franz Ebner, and Aljoscha C. Neubauer. To create or to recall? Neural mechanisms underlying the generation of creative new ideas. *NeuroImage*, 88:125–133, 2014.

[BJS13] Christer Bäckström, Peter Jonsson, and Simon Ståhlberg. Fast detection of unsolvable planning instances using local consistency. In *Proceedings of the Sixth Annual Symposium on Combinatorial Search*, pages 29–37, Leavenworth, Washington, USA, 2013.

[BKG09] Aron K Barbey, Frank Krueger, and Jordan Grafman. Structured event complexes in the medial prefrontal cortex support counterfactual representations for future planning. pages 1291–1300, 2009.

[BKR18] Jelle Bruineberg, Julian Kiverstein, and Erik Rietveld. The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195(6):2417–2444, 2018.

[BLS17] Chitta Baral, Barry Lumpkin, and Matthias Scheutz. A high level language for human robot interaction. *Advances in Cognitive Systems*, 2017.

[Bod10] Margaret A Boden. The Turing test and artistic creativity. *Kybernetes*, 3:409–413, 2010.

[BPP+15] Maddalena Boccia, Laura Piccardi, Liana Palermo, Raffaella Nori, and Massimiliano Palmiero. Where do bright ideas occur in ourbrain? Meta-analytic evidence from neuroimaging studies of domain-specific creativity. *Frontiers in Psychology*, 6(AUG):1–12, 2015.

[BR83] Mark H Bickhard and D Michael Richie. *On the nature of representation*. Citeseer, 1983.

[BR12] Stefan Baumann and Arndt Riester. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and meaning*, 25:119–162, 2012.

[Bro80] Gretchen P Brown. Characterizing indirect speech acts. *Computational Linguistics*, 6(3-4):150–166, 1980.

[BS13] Gordon Michael Briggs and Matthias Scheutz. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *AAAI*, 2013.

[BS16a] Gordon Briggs and Matthias Scheutz. The pragmatic social robot: Toward socially-sensitive utterance generation in human-robot interactions. In *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*, pages 12–15, 2016.

[BS16b] Selmer Bringsjord and Atriya Sen. On creative self-driving cars: Hire the computational logicians, fast. *Applied Artificial Intelligence*, 8:758–786, 2016.

[BSC02] Lawrence W Barsalou, Steven A Sloman, and Sergio E Chaigneau. The HIPE Theory of Function. In *Representing functional features for language and space: Insights from perception, categorization and development*, volume 30322, pages 1–25. 2002.

[BSH18] Sarah Brown-Schmidt and Daphna Heller. Perspective-taking during conversation. *Oxford handbook of psycholinguistics*, 2018.

[BWS17] Gordon Briggs, Tom Williams, and Matthias Scheutz. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 6(1):64–94, 2017.

[BWSH14] Marie-luise Brandi, Afra Wohlschla, Christian Sorg, and Joachim Hermsdo. The Neural Correlates of Planning and Executing Actual Tool Use. 34(39):13183–13194, 2014.

[BWTS17] Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ Inter-*

national Conference on Intelligent Robots and Systems (IROS), pages 6589–6594. IEEE, 2017.

[CACC09] Marco Campenní, Giulia Andrighetto, Federico Cecconi, and Rosaria Conte. Normal= normative? the role of intelligent agents in norm innovation. *Mind & Society*, 8(2):153, 2009.

[Cal10] Ryan Calo. Robots and privacy. robot ethics: The ethical and social implications of robotics, patrick lin, george bekey, and keith abney, eds, 2010.

[CANB13] N. Criado, E. Argente, P. Noriega, and V. Botti. Human-inspired model for norm compliance decision making. *Information Sciences*, 245:218–239, 2013.

[Cas05] Stephen Cass. Apollo 13, we have a solution. *IEEE Spectrum On-line, 04*, 1, 2005.

[CBCH16] Thomas R Colin, Tony Belpaeme, Angelo Cangelosi, and Nikolas Hemion. Hierarchical reinforcement learning as creative problem solving. *Robotics and Autonomous Systems*, 86:196–206, 2016.

[CBS05] Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1281–1288, Vancouver, Canada, 2005. MIT Press.

[CC94] James S Coleman and James Samuel Coleman. *Foundations of social theory*. Harvard university press, 1994.

[CFL+15] Michael Cashmore, Maria Fox, Derek Long, Daniele Magazzeni, Bram Ridder, Arnau Carrera, Narcis Palomeras, Natalia Hurtos, and Marc Carreras. Rosplan: Planning in the robot operating system. In *Twenty-Fifth International Conference on Automated Planning and Scheduling*, 2015.

[Che03] Anthony Chemero. An Outline of a Theory of Affordances. *Ecological Psychology*, 15(2):181–195, 2003.

[Che11] Anthony Chemero. *Radical embodied cognitive science*. MIT press, 2011.

[CKTK14] Tiffany L Chen, Chih-Hung Aaron King, Andrea L Thomaz, and Charles C Kemp. An investigation of responses to robot-initiated touch in a nursing context. *International Journal of Social Robotics*, 6(1):141–161, 2014.

[CM81] Herbert H Clark and Catherine R Marshall. Definite knowledge and mutual knowledge. *Elements of Discourse Understanding*, 1981.

[CM02] Herbert H Clark and Catherine R Marshall. Definite reference and mutual knowledge. *Psycholinguistics: critical concepts in psychology*, 414, 2002.

[CM11] Yun Chu and James N Macgregor. Human Performance on Insight Problem Solving : A Review Abstract : Keyword : 1 . Introduction : What Is Insight ? 3(2):119–150, 2011.

[CM16] Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, 2016.

[CMOS16] Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. A Bayesian approach to norm identification. *Frontiers in Artificial Intelligence and Applications*, 285:622–629, 2016.

[Coh05] Paul R Cohen. If not Turing's test, then what? *AI Magazine*, 26:61–67, 2005.

[Coh19] Philip R Cohen. Back to the future for dialogue research: A position paper. In *The Second AAAI Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL 2019), at the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019. Keynote Talk.

[CPCI15] Wesley P. Chan, Matthew K.X.J. Pan, Elizabeth A. Croft, and Masayuki Inaba. Characterization of handover orientations used by humans for efficient robot to human handovers. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE, sep 2015.

[CPQ06] Joyce Yue Chai, Zahar Prasov, and Shaolin Qu. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.

[CTS+12] Rehj Cantrell, Kartik Talamadupula, Paul Schermerhorn, J Benton, Subbarao Kambhampati, and Matthias Scheutz. Tell Me When and Why to Do It!: Run-time Planner Model Updates via Natural Language Instruction. In *Proceedings of the 2012 Human-Robot Interaction Conference*, 2012.

[CVL13] S Barry Cooper and Jan Van Leeuwen. *Alan Turing: His work and impact*. Elsevier, Waltham, MA, 2013.

[CW14] Hyun Jin Chung and Lisa L Weyandt. The Physiology of Executive Functioning. pages 13–28, 2014.

[Dar16] Kate Darling. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. *Robot Law, Calo, Froomkin, Kerr eds., Edward Elgar*, 2016.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Den17] Daniel C Dennett. *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company, 2017.

[dG16] Maartje MA de Graaf. An ethical evaluation of human–robot relationships. *International journal of social robotics*, 8(4):589–598, 2016.

[DHW+13] Tong Dandan, Zhu Haixue, Li Wenfu, Yang Wenjing, Qiu Jiang, and Zhang Qinglin. Brain activity in using heuristic prototype to solve insightful problems. *Behavioural Brain Research*, 253:139–144, 2013.

[DR12] Jan P De Ruiter. *Questions: Formal, functional and interactional perspectives*, volume 12. Cambridge University Press, 2012.

[DRC12] JP De Ruiter and Chris Cummins. A model of intentional communication: Airbus (asymmetric intention recognition with bayesian updating of signals). *Proceedings of SemDial 2012*, pages 149–50, 2012.

[dRng] J.P. de Ruiter. Turn-taking. forthcoming.

[DSBS09] Juraj Dzifcak, Matthias Scheutz, Chitta Baral, and Paul Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4163–4168, Kobe, Japan, 2009. IEEE.

[Dun45] Karl Duncker. On problem-solving. *Psychological Monographs*, 58(5):i–113, 1945.

[DWÖ16] Amory H Danek, Jennifer Wiley, and Michael Öllinger. Solving classical insight problems without aha! experience: 9 dot, 8 coin, and

matchstick arithmetic problems. *The Journal of Problem Solving*, 9(1):4, 2016.

[EGM97] Thomas Eiter, Georg Gottlob, and Heikki Mannila. Disjunctive datalog. *ACM Transactions on Database Systems (TODS)*, 22(3):364–418, 1997.

[EM17] Mihail Eric and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. *CoRR*, abs/1705.05414, 2017.

[ES13] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

[Fei03] Edward A. Feigenbaum. Some challenges and grand challenges for computational intelligence. *Journal of the ACM*, 50:32–40, 2003.

[Fes18] Leah Fessler. Amazon's alexa is now a feminist, and she's sorry if that upsets you, Jan 2018.

[FH91] Ronald Fagin and Joseph Y Halpern. Uncertainty, belief, and probability. *Computational Intelligence*, 7(3):160–173, 1991.

[FH13] Ronald Fagin and Joseph Y Halpern. A new approach to updating beliefs. *arXiv preprint arXiv:1304.1119*, 2013.

[FL07] Wolfgang Faber and Nicola Leone. On the complexity of answer set programming with aggregates. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 97–109. Springer, 2007.

[Fle08] Jessica I Fleck. Working memory demands in insight versus analytic problem solving. *European Journal of Cognitive Psychology*, 20(1):139–176, 2008.

[FMTK16] Jason Fischer, John G. Mikhael, Joshua B. Tenenbaum, and Nancy Kanwisher. Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, page 201610344, 2016.

[FN17] Lily Frank and Sven Nyholm. Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial Intelligence and Law*, 25(3):305–323, 2017.

[For10] Nigel Foreman. Virtual reality in psychology. *Themes in Science and Technology Education*, 2(1-2):225–252, 2010.

[FP15] Julia Fellrath and Radek Ptak. The role of visual saliency for the allocation of attention: Evidence from spatial neglect and hemianopia. *Neuropsychologia*, 73:70–81, 2015.

[FWL+13] Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology*, 4:859, 2013.

[FWS92] Ronald A Finke, Thomas B Ward, and Steven M Smith. Creative cognition: Theory, research, and applications. 1992.

[FZZ+16] X Fang, Y Zhang, Y Zhou, L Cheng, J Li, Y Wang, K J Friston, and T Jiang. Resting-State Coupling between Core Regions within the Central-Executive and Salience Networks Contributes to Working Memory Performance. *Frontiers in Behavioral Neuroscience*, 10(February):1–11, 2016.

[Gab16] Liane Gabora. The neural basis and evolution of divergent and convergent thought. *arXiv preprint arXiv:1611.03609*, 2016.

[Gal18] Shaun Gallagher. New mechanisms and the enactivist concept of constitution. *The metaphysics of consciousness*, pages 207–220, 2018.

[Gar15] John Gardner. The many faces of the reasonable person. *Law Quarterly Review*, 131(Oct), 2015.

[GC99] Nancy Green and Sandra Carberry. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435, 1999.

[GCS17] Denis Golovin, Jens Claßen, and Christoph Schwering. Reasoning about conditional beliefs for the Winograd Schema Challenge. 2017.

[GCS19] Evana Gizzi, Mateo Guaman Castro, and Jivko Sinapov. Creative problem solving by robots using action primitive discovery. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 228–233. IEEE, 2019.

[GF16] Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

[GHB+10] S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt. Interacting in time and space: Investigating human-human and human-robot joint action. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pages 252–257, 2010.

[GHC98] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Sixth Workshop on Very Large Corpora*, 1998.

[GHZ93] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307, 1993.

[Gib79] J J Gibson. *The ecological approach to visual perception*, volume 39. 1979.

[Gil16] Kenneth J Gilhooly. Incubation and intuition in creative problem solving. *Frontiers in psychology*, 7, 2016.

[GK14] Michael Gelfond and Yulia Kahl. *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach*. Cambridge University Press, 2014.

[GKE+10] Moritz Göbelbecker, Thomas Keller, Patrick Eyerich, Michael Brenner, and Bernhard Nebel. Coming up with good excuses: What to do when no plan can be found. In *Proceedings of the International Conference on Automated Planning and Scheduling*, page 81–88, Toronto, Canada, 2010. AAAI Press.

[GKK+08] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, Max Ostrowski, Torsten Schaub, and Sven Thiele. Engineering an incremental ASP solver. In *Proceedings of the International Conference on Logic Programming*, pages 190–205. Springer, 2008.

[GKK+11] Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider. Potassco: The potsdam answer set solving collection. *Ai Communications*, 24(2):107–124, 2011.

[GL75] David Gordon and George Lakoff. Conversational postulates. *Syntax and semantics 3: Speech acts*, 1975.

[GL90a] Michael Gelfond and Vladimir Lifschitz. *Logic programs with classical negation*, pages 579–597. MIT Press, 1990.

[GL90b] Michael Gelfond and Vladimir Lifschitz. Logic programs with classical negation, logic programming, 1990.

[Gła15] Paweł Gładziejewski. Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*, 40(1):63–90, 2015.

[GLP17] Peter Gregory, Alan Lindsay, and Julie Porteous. Domain model acquisition with missing information and noisy data. In *Workshop on Knowledge Engineering for Planning and Scheduling (KEPS). The 27th International Conference on Automated Planning and Scheduling (ICAPS)*, 2017.

[GLPK17] Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Strips planning in infinite domains. *arXiv preprint arXiv:1701.00287*, 2017.

[GN04] Malik Ghallab and Dana Nau. *Automated Planning: Theory and Practice*. 2004.

[GN12] Gazzaley, Adam and Anna C. Nobre. Top-down modulation: Bridging selective attention and working memory. *Trends Cogn. Sci.*, 60(2):830–846, 2012.

[Gri75] H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

[GS86] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

[Gui62] Joy Paul Guilford. Creativity: Its measurement and development. *A source book for creative thinking*, pages 151–167, 1962.

[Gut12] Sinziana Gutiu. Sex robots and roboticization of consent. In *We Robot 2012 conference, Coral Gables, Florida. Retrieved April*, volume 15, page 2013, 2012.

[GVS14] Patricia Garrido-Vásquez and Anna Schubö. Modulation of Visual Attention by Object Affordance. *Frontiers in Psychology*, 5(February):1–11, 2014.

[GW12] Sabrina Golonka and Andrew D Wilson. Gibson's ecological approach. *Avant: Trends in Interdisciplinary Studies 3 (2)*, pages 40–53, 2012.

[HA89] Elizabeth A Hinkelman and James F Allen. Two constraints on speech act ambiguity. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 212–219. Association for Computational Linguistics, 1989.

[Hab15] Jürgen Habermas. *The Theory of Communicative Action: Lifeworld and Systems, a Critique of Functionalist Reason*, volume 2. John Wiley & Sons, 2015.

[Har90] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[Har91] Stevan Harnad. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1:43–54, 1991.

[Hay07] Ryan L. Hayes SM, Nadel L. The Effect of Scene Context on Episodic Object Recognition: Parahippocampal Cortex Mediates Memory Encoding and Retrieval Success. *Hippocampus*, 9(1):19–22, 2007.

[HBB+15] Aidan J Horner, James A Bisby, Daniel Bush, Wen-Jing Lin, and Neil Burgess. Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature communications*, 6(May):7462, 2015.

[HCL+13] Xin Hao, Shuai Cui, Wenfu Li, Wenjing Yang, Jiang Qiu, and Qinglin Zhang. Enhancing insight in scientific problem solving by highlighting the functional features of prototypes: An fMRI study. *Brain Research*, 1534:46–54, 2013.

[HdMdBW14] Andreas Herzig, Maria Viviane de Menezes, Leliane Nunes de Barros, and Renata Wassermann. On the revision of planning tasks. In *Proceedings of the Twenty-First European Conference on Artificial Intelligence*, pages 435–440, Prague, Czech Republic, 2014. EurAI.

[HEHEG19] Manuel Heras-Escribano, Manuel Heras-Escribano, and George. *The philosophy of affordances*. Springer, 2019.

[Her14] Andreas Herzig. Belief change operations: A short history of nearly everything, told in dynamic logic of propositional assignments. In *KR*, 2014.

[HFA+15] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. Robot presence and human honesty: Experimental evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 181–188. ACM, 2015.

[Hir81] Grahame Hirst. Anaphora in natural language understanding. 1981.

[HK17] Elliott M Hoey and Kobin H Kendrick. Conversation analysis. *Research methods in psycholinguistics: A practical guide*, pages 151–173, 2017.

[HLM+17] Chris Haynes, Michael Luck, Peter McBurney, Samhar Mahmoud, Tomáš Vítek, and Simon Miles. Engineering the emergence of norms: a review. *The Knowledge Engineering Review*, 32, 2017.

[HNH+16] Jarmo Heinonen, Jussi Numminen, Yevhen Hlushchuk, Henrik Antell, Vesa Taatila, and Jyrki Suomala. Default mode and executive networks areas: Association with the serial order in divergent thinking. *PLoS ONE*, 11(9):1–16, 2016.

[Hob78] Jerry R Hobbs. Resolving pronoun references. *Lingua*, 44(4):311–338, 1978.

[HRCS16] Xiaowei Huang, Ji Ruan, Qingliang Chen, and Kaile Su. Normative multiagent systems: a dynamic generalization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1123–1129. AAAI Press, 2016.

[HS17] Todd Hester and Peter Stone. Intrinsically motivated model learning for developing curious robots. *Artificial Intelligence*, 247:170–186, 2017.

[Hur96] Heidi M Hurd. The moral magic of consent. *Legal Theory*, 2(2):121–146, 1996.

[IDN87] Alice M. Isen, Kimberly a. Daubman, and Gary P. Nowicki. Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52(6):1122–1131, 1987.

[IJVE16] Marcello Ienca, Fabrice Jotterand, Constantin Vică, and Bernice Elger. Social and assistive robotics in dementia care: ethical recommendations for research and practice. *International Journal of Social Robotics*, 8(4):565–573, 2016.

[IVS18] Filip Ilievski, Piek Vossen, and Stefan Schlobach. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[JBC+98] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Johns Hopkins LVCSR Workshop-97, Switchboard discourse language modeling project, Final Report. 1998.

[JBN12] Emanuel Jauk, Mathias Benedek, and Aljoscha C. Neubauer. Tackling creativity at its roots: Evidence for different patterns of EEG alpha activity related to convergent and divergent modes of task processing. *International Journal of Psychophysiology*, 84(2):219–225, 2012.

[JCL+20] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[JHvdM+17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1997, Honolulu, HI, 2017.

[JM09] Kristiina Jokinen and Michael McTear. *Spoken dialogue systems*. Morgan & Claypool Publishers, 2009.

[JMH15] Malte F Jung, Nikolas Martelaro, and Pamela J Hinds. Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 229–236. ACM, 2015.

[JZMMUL18] Salud María Jiménez-Zafra, Roser Morante, Maite Martin, and L. Alfonso Ureña-López. A review of Spanish corpora annotated with negation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 915–924, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[KAAM16] Yoed N. Kenett, M. M. Allaham, Joseph L. Austerweil, and Bertram F. Malle. The norm fluency task: Unveiling the properties of norm representation. (Poster.). In *Poster presented at the 57th*

Annual Meeting of the Psychonomic Society, Boston, MA, November 2016. November 2016.

[Kam81] Hans Kamp. A theory of truth and semantic representation. *Formal semantics-the essential readings*, pages 189–222, 1981.

[Kan08] Immanuel Kant. Critique of pure reason. 1781. *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, pages 370–456, 1908.

[KAS18] Daniel Kasenberg, Thomas Arnold, and Matthias Scheutz. Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 1st AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2018.

[Kau11] Scott Barry Kaufman. Intelligence and the cognitive unconscious. *The Cambridge handbook of intelligence*, pages 442–467, 2011.

[KB09] George Konidaris and Andrew G. Barto. Skill Discovery in Continuous Reinforcement Learning Domains using Skill Chaining. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1015–1023. Curran Associates, Inc., 2009.

[KB14] John Kounios and Mark Beeman. The cognitive neuroscience of insight. *Annual review of psychology*, 65, 2014.

[KC15] Fadhela Kerdjoudj and Olivier Curé. Evaluating uncertainty in textual document. In *Proceedings of the Eleventh International Workshop on Uncertainty Reasoning for the Semantic Web*, page 1, Bethlehem, PA, USA, 2015. Springer.

[Keh00] Andrew Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the 17th AAAI Conference on Artificial Intelligence*, pages 685–690, 2000.

[Key07] Boaz Keysar. Communication and miscommunication: The role of egocentric processes, 2007.

[KHM16] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

[KJKI+15] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Solace Shen, Heather E Gary, and Jolina H Ruckert. Will people keep the secret of a humanoid robot?: Psychological intimacy in HRI. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 173–180. ACM, 2015.

[KKLP14] George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. Constructing symbolic representations for high-level planning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1932–1938, Quebec City, Canada, 2014.

[KKLP18] George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, 61:215–289, 2018.

[KLP13] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, 32:1194–1227, 2013.

[KM17] Guilherme Krzisch and Felipe Meneguzzi. Norm Identification in Jason using a Bayesian Approach. In *Proceedings of the AAMAS Workshop of Multi-Agent-Based Simulation*, 2017.

[Kno09] Günther Knoblich. Psychological research on insight problem solving. In Hans Primas Harald Atmanspacher, editor, *Recasting reality*, pages 275–300. Springer, 2009.

[KP04] Jinsul Kim and Jihwan Park. Advanced Grasp Planning for Handover Operation Between Human and Robot: Three Handover Methods in Esteem Etiquettes Using Dual Arms and Hands of Home-Service Robot. *2nd International Conference on Autonomous Robots and Agents*, (c):34–39, 2004.

[KPD+04] EC Kulasekere, Kamal Premaratne, Duminda A Dewasurendra, M-L Shyu, and Peter H Bauer. Conditioning and updating evidence. *International Journal of Approximate Reasoning*, 36(1):75–108, 2004.

[Kri18] Beate Krickel. Saving the mutual manipulability account of constitutive relevance. *Studies in History and Philosophy of Science Part A*, 68:58–67, 2018.

[KS89] Robert Kowalski and Marek Sergot. A logic-based calculus of events. In *Foundations of knowledge base management*, pages 23–55. Springer, 1989.

[KS90] Craig A Kaplan and Herbert A Simon. In search of insight. *Cognitive psychology*, 22(3):374–419, 1990.

[KS16a] Radimir Komarnitsky and Guy Shani. Computing contingent plans using online replanning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[KS16b] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, (1):14–29, 2016.

[KSAO+14] Kheng Lee Koay, Dag Sverre Syrdal, Mohammadreza Ashgari-Oskoei, Michael L Walters, and Kerstin Dautenhahn. Social roles and baseline proxemic preferences for a domestic service robot. *International Journal of Social Robotics*, 6(4):469–488, 2014.

[KSTN11] Marc Kammer, Thomas Schack, Marko Tscherepanow, and Yukie Nagai. From affordances to situated affordances in robotics-why context is important. In *Frontiers in Computational Neuroscience*, number 30, 2011.

[Lai12] John E Laird. *The Soar cognitive architecture*. MIT press, 2012.

[Lan17] Pat Langley. Heuristics and cognitive systems. *Advances in Cognitive Systems*, 5:3–12, 2017.

[LDM12a] Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[LDM12b] Hector J Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561. AAAI Press, 2012.

[Leu12] Bert Leuridan. What are mechanisms in social science?, 2012.

[Lev01] Stephen C Levinson. Pragmatics. In *International Encyclopedia of Social and Behavioral Sciences: Vol. 17*, pages 11948–11954. Pergamon, 2001.

[LHTL17] Yang Liu, Kun Han, Zhao Tan, and Yun Lei. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[LJ88] Pat Langley and Randolph Jones. A computational model of scientific insight. In Robert J. Sternberg, editor, *The Nature of Creativity: Contemporary Psychological Perspectives*, pages 177–201. Cambridge University Press, New York, NY, 1988.

[LJGDR15] Sebastian Loth, Katharina Jettka, Manuel Giuliani, and Jan P De Ruiter. Ghost-in-the-machine reveals human social signals for human–robot interaction. *Frontiers in psychology*, 6, 2015.

[LJL+16] Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in Winograd Schema Challenge. *arXiv preprint arXiv:1611.04146*, 2016.

[LK06] Jim Lovell and Jeffrey Kluger. *Apollo 13*. Houghton Mifflin Harcourt, 2006.

[LL94] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561, 1994.

[LLC+19] Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China, November 2019. Association for Computational Linguistics.

[LLQ+13] Junlong Luo, Wenfu Li, Jiang Qiu, Dongtao Wei, Yijun Liu, and Qinlin Zhang. Neural Basis of Scientific Innovation Induced by Heuristic Prototype. *PLoS ONE*, 8(1):1–7, 2013.

[LMG16] Nir Lipovetzky, Christian J Muise, and Hector Geffner. Traps, invariants, and dead-ends. In *Proceedings of the International Conference*

on Automated Planning and Scheduling, pages 211–215, London, UK, 2016. AAAI Press.

[LNR87] John E Laird, Allen Newell, and Paul S Rosenbloom. Soar: An architecture for general intelligence. *Artificial intelligence*, (1):1–64, 1987.

[LPB+16] Pat Langley, Chris Pearce, Yu Bai, M Barley, and C Worsfold. Variations on a theory of problem solving. In *Proceedings of the Fourth Annual Conference on Cognitive Systems*, Evanston, IL, 2016.

[LRF+17] Alan Lindsay, Jonathon Read, Joao F Ferreira, Thomas Hayton, Julie Porteous, and Peter Gregory. Framer: Planning models from natural language action descriptions. In *Twenty-Seventh International Conference on Automated Planning and Scheduling*, 2017.

[LSAJJ16] Damir Lotinac, Javier Segovia-Aguas, Sergio Jiménez, and Anders Jonsson. Automatic generation of high-level state features for generalized planning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence; 2016 July 9-15; New York, United States. Palo Alto: AAAI Press; 2016. p. 3199-3205.* Association for the Advancement of Artificial Intelligence (AAAI), 2016.

[LSS+17] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. *CoRR*, abs/1710.03957, 2017.

[LW20] Jane Lockshin and Tom Williams. We need to start thinking ahead: The impact of social context on linguistic norm adherence. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 2020.

[Mai30] Norman RF Maier. Reasoning in humans. i. on direction. *Journal of comparative Psychology*, 10(2):115, 1930.

[Mar77] Edna Ullman Margalit. *The emergence of norms*. Oxford, 1977.

[MB14] Laura Macchi and Maria Bagassi. The interpretative heuristic in insight problem solving. *Mind & Society*, 13(1):97–108, 2014.

[MBS19] Bertram F Malle, Paul Bello, and Matthias Scheutz. Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 21–27, 2019.

[MCR+18] Marcin Miłkowski, Robert Clowes, Zuzanna Rucińska, Aleksandra Przegalińska, Tadeusz Zawidzki, Joel Krueger, Adam Gies, Marek McGann, Łukasz Afeltowicz, Witold Wachowski, et al. From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology*, 9:2393, 2018.

[MDO16] Leora Morgenstern, Ernest Davis, and Charles L Ortiz. Planning, executing, and evaluating the Winograd Schema Challenge. *AI Magazine*, 37(1):50–54, 2016.

[Men76] Gerald Mendelsohn. Associative and attentional processes in creative performance. *Journal of Personality*, 44:341–369, 1976.

[Men15] V. Menon. *Salience Network*, volume 2. Elsevier Inc., 2015.

[Met86] Janet Metcalfe. Premonitions of insight predict impending error. *Journal of experimental psychology: Learning, memory, and cognition*, 12(4):623, 1986.

[MFE+00] A Miyake, N P Friedman, M J Emerson, a H Witzki, A Howerter, and T D Wager. The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive psychology*, 41(1):49–100, 2000.

[Mit14] Ruslan Mitkov. *Anaphora resolution*. Routledge, 2014.

[MJ13] Robert A. Mason and Marcel Adam Just. Neural representations of physics concepts. *Psychological Science*, pages 1–9, 2013.

[ML09] L. Montesano and M. Lopes. Learning grasping affordances from local visual descriptors. *2009 IEEE 8th International Conference on Development and Learning (ICDL 2009)*, 2009.

[MLBSV07] Luis Montesano, Manuel Lop, Alexandre Bernardino, and Jose Santos-Victor. Modeling affordances using Bayesian networks. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4102–4107, 2007.

[MLSRA+14] Javier Morales, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Michael Wooldridge, and Wamberto Vasconcelos. Minimality and Simplicity in the On-line Automated Synthesis of Normative Systems. *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, pages 109–116, 2014.

[MM11] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 331–338. ACM, 2011.

[MM15] Ross Mead and Maja J Mataric. Robots have needs too: People adapt their proxemic preferences to improve autonomous robot recognition of human social signals. *New Frontiers in Human-Robot Interaction*, page 100, 2015.

[MM17] Ross Mead and Maja J Matarić. Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *Autonomous Robots*, 41(5):1189–1201, 2017.

[MMM+12] JV McDonnell, JB Martin, DB Markant, A Coenen, AS Rich, and TM Gureckis. psiturk (version 1.02)[software]. new york, ny: New york university, 2012.

[MMV+12] Bogdan Moldovan, Plinio Moreno, Martijn Van Otterlo, José Santos-Victor, and Luc De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 4373–4378, 2012.

[MOC01] James N MacGregor, Thomas C Ormerod, and Edward P Chronicle. Information processing and insight: a process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):176, 2001.

[MP76] David Marr and Tomaso Poggio. From understanding computation to understanding neural circuitry. 1976.

[MS12] Fulvio Mastrogiovanni and Antonio Sgorbissa. A biologically plausible, neural-inspired planning approach which does not solve 'The gourd, the monkey, and the rice' puzzle. *Biologically Inspired Cognitive Architectures*, 2:77–87, 2012.

[MS15] Bertram F Malle and Matthias Scheutz. When will people regard robots as morally competent social partners? In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 486–491. IEEE, 2015.

[MSA17a] Bertram F. Malle, Matthias Scheutz, and Joseph L. Austerweil. Networks of social and moral norms in human and robot agents. In Maria Isabel Aldinhas Ferreira, Joao Silva Sequeira, Mohammad Osman Tokhi, Endre E. Kadar, and Gurvinder Singh Virk, editors, *A World with Robots*, pages 3–17. Springer, Cham, Switzerland, 2017.

[MSA17b] BF Malle, M Scheutz, and JL Austerweil. Networks of social and moral norms in human and robot agents. In *A World with Robots*, pages 3–17. Springer, 2017.

[MSSZ11] Fulvio Mastrogiovanni, Antonello Scalmato, Antonio Sgorbissa, and Renato Zaccaria. Problem Awareness for Skilled Humanoid Robots. *International Journal of Machine Consciousness*, 3(1):91–114, 2011.

[MTFA15] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1374–1381, Seattle, WA, USA, 2015. IEEE.

[MZ09] Ravi Mehta and Rui (Juliet) Zhu. Blue or red? exploring the effect of color on cognitive task performances. *Science*, 323(5918):1226–1229, 2009.

[MZPS12] Kira Mourao, Luke S Zettlemoyer, Ronald Petrick, and Mark Steedman. Learning STRIPS operators from noisy and incomplete observations. *arXiv preprint arXiv:1210.4889*, 2012.

[NCJH19] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in minecraft. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5405–5415, 2019.

[NDRA10] Jason Noble, Jan Peter De Ruiter, and Kate Arnold. From monkey alarm calls to human language: How simulations can fill the gap. *Adaptive Behavior*, 18(1):66–82, 2010.

[NDS+13] R C Nunez, R Dabarera, M Scheutz, G Briggs, O Bueno, K Premaratne, and M N Murthi. DS-based uncertain implication rules for inference and fusion applications. *Information Fusion (FUSION), 2013 16th International Conference on*, pages 1934–1941, 2013.

[Ng10] Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 1396–1411. Association for Computational Linguistics, 2010.

[NG15] S D Newman and S R Green. Complex Problem Solving. *Brain Mapping: An Encyclopedic Reference*, 3:543–549, 2015.

[NI05] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, 2005.

[NN+04] Alva Noë, Alva Noë, et al. *Action in perception*. MIT press, 2004.

[Nor88] Don Norman. The Psychology of Everyday Things. In *The Psychology of Everyday Things*, pages 1–104. 1988.

[NPL+20] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey, 2020.

[NR15] Aastha Nigam and Laurel D Riek. Social context perception for mobile robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3621–3627, Hamburg, Germany, 2015. IEEE.

[NSZC19] Lakshmi Nair, Nithin Shrivastav, Erickson Zackory, and Sonia Chernova. Autonomous tool construction using part shape and attachment prediction. In *Proceedings of Robotics: Science and Systems*, 2019.

[OA14] Billy Okal and Kai O Arras. Towards group-level social activity recognition for mobile robots. In *Proceedings of the Workshop on Assistance and Service Robotics in a Human Environments Workshop at the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Chicago, IL, USA, 2014. IEEE.

[OB16] Joachim T. Operskalski and Aron K. Barbey. Cognitive Neuroscience of Causal Reasoning. *Oxford Handbook of Causal Reasoning*, 2016.

[OED19] OED Online. schema, n., 2019. https://www.oed.com/ view/Entry/172307 (accessed December 02, 2019).

[OF14] Ana-Maria Olteteanu and Christian Freksa. Towards affordance-based solving of object insight problems. In *Proceedings of First Workshop on Affordances: Affordances in Vision for Cognitive Robotics, Robotics Science and Systems*, 2014.

[ÖFBS17] Michael Öllinger, Anna Fedor, Svenja Brodt, and Eörs Szathmáry. Insight into the ten-penny problem: guiding search by constraints and maximization. *Psychological research*, 81(5):925–938, 2017.

[Ohl92] Stellan Ohlsson. Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking*, 1:1–44, 1992.

[ÖJK14] Michael Öllinger, Gary Jones, and Günther Knoblich. The dynamics of search, impasse, and representational change provide a coherent explanation of difficulty in the nine-dot problem. *Psychological research*, 78(2):266–275, 2014.

[OM13] Nir Oren and Felipe Meneguzzi. Norm identification through plan recognition. In *Proceedings of the workshop on Coordination, Organization, Institutions and Norms in Agent Systems (COIN 2013@ AAMAS)*, 2013.

[OR15] Maria Francesca O'Connor and Laurel D Riek. Detecting social context: A method for social event classification using naturalistic multimodal data. In *Proceedings of the Eleventh IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–7, Ljubljana, Slovenia, 2015. IEEE.

[PA80] C Raymond Perrault and James F Allen. A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6(3-4):167–182, 1980.

[PAED17] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the Thirty-fourth International Conference on Machine Learning*, pages 2778–2787, Sydney, Australia, 2017. PMLR.

[PFS+15] Priyam Parashar, Robert Fisher, Reid Simmons, Manuela Veloso, and Joydeep Biswas. Learning context-based outcomes for mobile robots in unstructured indoor environments. In *Proceedings of the Fourteenth IEEE International Conference on Machine Learning and Applications*, pages 703–706, Miami, FL, USA, 2015. IEEE.

[PKR15] Haoruo Peng, Daniel Khashabi, and Dan Roth. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, 2015.

[PMZ+09] Kamal Premaratne, Manohar N Murthi, Jinsong Zhang, Matthias Scheutz, and Peter H Bauer. A dempster-shafer theoretic conditional approach to evidence updating for fusion of hard and soft data. In *2009 12th International Conference on Information Fusion*, pages 2122–2129. IEEE, 2009.

[PNŠ+17] Jennifer Piatt, Shinichi Nagata, Selma Šabanović, Wan-Ling Cheng, Casey Bennett, MS Hee Rin Lee, David Hakken, et al. Companionship with a robot? therapists' perspectives on socially assistive robots as therapeutic interventions in community mental health for older adults. *American Journal of Recreation Therapy*, 15(4):29–39, 2017.

[POYI18] Oskar Palinko, Kohei Ogawa, Yuichiro Yoshikawa, and Hiroshi Ishiguro. How should a robot interrupt a conversation between multiple

humans. In *International Conference on Social Robotics*, pages 149–159. Springer, 2018.

[PPMS16] Lalintha G Polpitiya, Kamal Premaratne, Manohar N Murthi, and Dilip Sarkar. A framework for efficient computation of belief theoretic operations. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 1570–1577. IEEE, 2016.

[PPMS17] Lalintha G Polpitiya, Kamal Premaratne, Manohar N Murthi, and Dilip Sarkar. Efficient computation of belief theoretic conditionals. In *Proceedings of the tenth international symposium on imprecise probability: Theories and applications*, pages 265–276, 2017.

[PRZV14] Ekaterina Potapova, Andreas Richtsfeld, Michael Zillich, and Markus Vincze. Incremental Attention-driven Object Segmentation. In *14th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 252—-258, 2014.

[PS09] Luís Moniz Pereira and Ari Saptawijaya. Modelling morality with prospective logic. *International Journal of Reasoning-based Intelligent Systems*, 1(3-4):209–221, 2009.

[QCG+09] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3. Kobe, Japan, 2009.

[QRP08] René Quilodran, Marie Rothé, and Emmanuel Procyk. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, 57(2):314–325, 2008.

[RBG11] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

[RBGAC17] Adam Richard-Bollans, L Gomez Alvarez, and Anthony G Cohn. The role of pragmatics in solving the Winograd Schema Challenge. In *Proceedings of 13th International Symposium on Commonsense Reasoning (Commonsense-2017)*. CEUR Workshop Proceedings, 2017.

[RD14] Simone M Ritter and Ap Dijksterhuis. Creativity—the unconscious foundations of the incubation period. *Frontiers in human neuroscience*, 8, 2014.

[RDF15] M Birna Van Riemsdijk, Louise Dennis, and Michael Fisher. A Semantic Framework for Socially Adaptive Agents Towards Strong Norm Compliance. *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, pages 423–432, 2015.

[RDT15] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

[Ree96] Edward.s Reed. *Encountering the world: toward an ecological psychology*, volume 34. 1996.

[Rei19] Ehud Reiter. Natural language generation challenges for explainable ai. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7, 2019.

[RH11] Katherine L Roberts and Glyn W Humphreys. Action-related objects influence the distribution of visuospatial attention. *Quarterly journal of experimental psychology (2006)*, 64(January 2012):669–688, 2011.

[Rie14] Mark O. Riedl. The Lovelace 2.0 test of artificial creativity and intelligence. *arXiv preprint arXiv:1410.6142*, 2014.

[RJ19] Edward Reed and Rebecca Jones. *Reasons for realism: Selected essays of James J. Gibson*. Routledge, 2019.

[RK14] Erik Rietveld and Julian Kiverstein. A Rich Landscape of Affordances. *Ecological Psychology*, 26(4):325–352, 2014.

[RMMG08] Vasile Rus, Philip M McCarthy, Danielle S McNamara, and Arthur C Graesser. A study of textual entailment. *International Journal on Artificial Intelligence Tools*, 17(04):659–685, 2008.

[RN12] Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics, 2012.

[Rob16] SI Robertson. *Problem Solving: Perspectives from Cognition and Neuroscience*. Psychology Press, 2016.

[RR11] Laurel D Riek and Peter Robinson. Challenges and opportunities in building socially intelligent machines [social sciences]. *IEEE Signal Processing Magazine*, 28(3):146–149, 2011.

[RTO19] Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1488, 2019.

[RTSS20] Antonio Roque, Alexander Tsuetaki, Vasanth Sarathy, and Matthias Scheutz. Developing a corpus of indirect speech act schemas. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2020.

[Ruc17] Zuzanna Rucińska. The role of affordances in pretend play. *Embodiment, enaction, and culture: Investigating the constitution of the shared world*, pages 257–278, 2017.

[Ruc20] Zuzanna Rucińska. Affordances in dennett's from bacteria to bach and back, 2020.

[SA07] Sandip Sen and Stephane Airiau. Emergence of Norms Through Social Learning. *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, pages 1507–1512, 2007.

[SA16] Matthias Scheutz and Thomas Arnold. Are we ready for sex robots? In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction*, 2016.

[SA17] Matthias Scheutz and Thomas Arnold. Intimacy, bonding, and sex robots: Examining empirical results and exploring ethical ramifications. In John Danaher and Neil McArthur, editors, *Robot Sex: Social and Ethical Implications*. MIT Press, 2017.

[SAFMJ16] Javier Segovia-Aguas, Jonathan Ferrer-Mestres, and Anders Jonsson. Planning with partially specified behaviors. In *Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence*, volume 288, 2016.

[Sar18a] Vasanth Sarathy. Real world problem-solving. *Frontiers in Human Neuroscience*, 12:1–14, 2018.

[Sar18b] Vasanth Sarathy. Real world problem-solving. *Frontiers in human neuroscience*, 12, 2018.

[Sar19] Vasanth Sarathy. Learning context-sensitive norms under uncertainty. In *Proceedings of the 2nd AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES-19)*, 2019.

[SAS19] Vasanth Sarathy, Thomas Arnold, and Matthias Scheutz. When exceptions are the norm: Exploring the role of consent in hri. *ACM Transactions on Human-Robot Interaction (Formerly, Journal of Human-Robot Interaction)*, 8(3):14:1–14:21, July 2019.

[Saw11] Keith Sawyer. The Cognitive Neuroscience of Creativity: A Critical Review. *Creativity Research Journal*, 23(2):137–154, 2011.

[SB12] Jason M. Scimeca and David Badre. Striatal contributions to declarative memory retrieval Jason. *Neuron*, 75(3):380–392, 2012.

[SB16] Carola Salvi and Edward M Bowden. Looking for creativity: Where do we look when we look for new ideas? *Frontiers in psychology*, 7, 2016.

[SBBC19] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WINOGRANDE: An adversarial Winograd Schema Challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

[SBC+13] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proceedings of the AAAI Workshop on Intelligent Robotic Systems*, Bellevue, WA, USA, 2013. AAAI Press.

[SC99] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.

[SC11] Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm Creation, Spreading and Emergence: A Survey of Simulation Models of Norms in Multi-Agent Systems. *Multiagent and Grid Systems*, 7(1):21–54, 2011.

[Sca03] Andrea Scarantino. Affordances Explained. *Philosophy of Science*, 70(5):949–961, 2003.

[SCD+07] E. Sahin, M. Cakmak, M. R. Dogar, E. Ugur, and G. Ucoluk. To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control. *Adaptive Behavior*, 15(4):447–472, 2007.

[Sch07] R. C. Schmidt. Scaffolds for Social Meaning. *Ecological Psychology*, 19(2):137–151, 2007.

[Sch11] Matthias Scheutz. 13 the inherent dangers of unidirectional emotional bonds between humans and social robots. *Robot ethics: The ethical and social implications of robotics*, page 205, 2011.

[Sch12] Paul Schweizer. The externalist foundations of a truly total turing test. *Minds and Machines*, 22:191–212, 2012.

[Sch14] Peter Schüller. Tackling Winograd Schemas by formalizing relevance theory in knowledge graphs. In *Proceedings of the Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2014.

[Sch17] Matthias Scheutz. The case for explicit ethical agents. *AI Magazine*, 38(4):57–64, 2017.

[SCPP13] Bastin Tony Roy Savarimuthu, Stephen Cranefield, Maryam A. Purvis, and Martin K. Purvis. Identifying prohibition norms in agent societies. *Artificial Intelligence and Law*, 21(1):1–46, 2013.

[SCS14] Megan Strait, Cody Canning, and Matthias Scheutz. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 479–486. ACM, 2014.

[SDB+04] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, 2004.

[Sea75] John R Searle. Indirect speech acts. *Syntax & Semantics, 3: Speech Act*, pages 59–82, 1975.

[SFK+06] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40. ACM, 2006.

[SFR+14] Siddharth Srivastava, Eugene Fang, Lorenzo Riano, Rohan Chitnis, Stuart Russell, and Pieter Abbeel. Combined task and motion planning through an extensible planner-independent interface layer. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 639–646, Hong Kong, China, 2014. IEEE.

[SGC19] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

[SGGC07] Ja-Young Sung, Lan Guo, Rebecca E Grinter, and Henrik I Christensen. "my roomba is rambo": intimate home appliances. In *International Conference on Ubiquitous Computing*, pages 145–162. Springer, 2007.

[Sha76] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[SHM10] Jennifer J Summerfield, Demis Hassabis, and Eleanor A Maguire. Differential engagement of brain regions within a 'core'network during scene construction. *Neuropsychologia*, 48(5):1501–1509, 2010.

[Sim87] Kenneth W Simons. Assumption of risk and consent in the law of torts: A theory of full preference. *BUL Rev.*, 67:213, 1987.

[Sim06] Kenneth W Simons. A restatement (third) of intentional torts. *Ariz. L. Rev.*, 48:1061, 2006.

[Sim08] Joydeep Bhattacharya Simone Sandkühler. Deconstructing Insight: EEG Correlates of Insightful Problem Solving. *PLoS One*, 101(11):1435–1439, 2008.

[Sim17] Kenneth Simons. Actual, apparent and hypothetical consent in tort law, 2017. https://www.law.uci.edu/centers/clp/images-pdfs/simons-actual-apparent-and-hypothetical-consent.pdf.

[SJWE12] Aziez Sardar, Michiel Joosse, Astrid Weiss, and Vanessa Evers. Don't stand so close to me: users' attitudinal and behavioral responses to personal space invasion by robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 229–230. ACM, 2012.

[SKCT06] Aaron Sloman, Geert-Jan Kruijff, Jackie Chappell, and Arnold Trehub. Sensorimotor vs objective contingencies. *University of Bimringham, School of Computer Science, Tech. Rep. COSY-DP-0603*, 2006.

[SKO+17] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the Sixteenth International Conference on Autonomous Agents and Multiagent Systems*, Sao Paolo, Brazil, 2017. IFAAMAS.

[SKWD07] Dag Sverre Syrdal, Kheng Lee Koay, Michael L Walters, and Kerstin Dautenhahn. A personalized robot companion?-the role of individual differences on spatial preferences in hri scenarios. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 1143–1148. IEEE, 2007.

[SLD+13] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, Siddhartha S Srinivasa, Maya Cakmak, Vincenzo Micelli, and Willow

Garage. Toward Seamless Human – Robot Handovers. *Journal of Human Robot Interaction*, 2(1):112–132, 2013.

[Slo11] Aaron Sloman. What's information, for an organism or intelligent machine? how can a machine or organism mean? In *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*, pages 393–438. World Scientific, 2011.

[SLOA12] Andrea Stocco, Christian Lebiere, Randall C O'Reilly, and John R Anderson. Distinct contributions of the caudate nucleus, rostral prefrontal cortex, and parietal cortex to the execution of instructed tasks. *Cognitive, affective & behavioral neuroscience*, 12(4):611–628, 2012.

[SM14] Matthias Scheutz and Bertram F. Malle. "Think and do the right thing"—A plea for morally competent autonomous robots. In *Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on*, pages 1–4. IEEE, 2014.

[SM17] Mohan Sridharan and Ben Meadows. Learning affordances for assistive robots. In *Proceedings of the International Conference on Social Robotics*, Tsukuba, Japan, 2017. Springer.

[SOKS18] Vasanth Sarathy, Bradley Oosterveld, Evan Krause, and Matthias Scheutz. Learning cognitive affordances for objects from natural language instruction. In *Proceedings of the Sixth Annual Conference on Advances in Cognitive Systems*, 2018.

[SPG15] Paul T Sowden, Andrew Pringle, and Liane Gabora. The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & Reasoning*, 21(1):40–60, 2015.

[SPS99] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[SRE+17] Giulia Sprugnoli, Simone Rossi, Alexandra Emmendorfer, Alessandro Rossi, Sook-Lei Liew, Elisa Tatti, Giorgio di Lorenzo, Alvaro Pascual-Leone, and Emiliano Santarnecchi. Neural correlates of Eureka moment. *Intelligence*, 2017.

[SRZ16] Tianmin Shu, M. S. Ryoo, and Song-Chun Zhu. Learning Social Affordance for Human-Robot Interaction. *International Joint Conference on Artificial Intelligence (IJCAI), 2016*, page Accepted, apr 2016.

[SS15a] Vasanth Sarathy and Matthias Scheutz. Semantic Representation of Objects and Function. In *Proceedings of the 2015 IROS Workshop on Learning Object Affordances*, 2015.

[SS15b] Vasanth Sarathy and Matthias Scheutz. Semantic representation of objects and function. In *Proceedings of the 2015 IROS Workshop on Learning Object Affordances*, 2015.

[SS16a] Vasanth Sarathy and Matthias Scheutz. Beyond grasping - perceiving affordances across various stages of cognitive development. In *Proceedings of the The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics (ICDL)*, 2016.

[SS16b] Vasanth Sarathy and Matthias Scheutz. Cognitive affordance representations in uncertain logic. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, Cape Town, South Africa, 2016. AAAI Press.

[SS16c] Vasanth Sarathy and Matthias Scheutz. Cognitive affordance representations in uncertain logic. In *Proceedings of the 15th International*

Conference on Principles of Knowledge Representation and Reasoning (KR), 2016.

[SS16d] Vasanth Sarathy and Matthias Scheutz. A logic-based computational framework for inferring cognitive affordances. *IEEE Transactions on Cognitive and Developmental Systems*, 8(3), 2016.

[SS17] Christopher John Stanton and Catherine J Stevens. Don't stare at me: The impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task. *International Journal of Social Robotics*, 9(5):745–753, 2017.

[SS18a] Vasanth Sarathy and Matthias Scheutz. A logic-based computational framework for inferring cognitive affordances. *IEEE Transactions on Cognitive and Developmental Systems*, (1):26–43, 2018.

[SS18b] Vasanth Sarathy and Matthias Scheutz. Macgyver problems: Ai challenges for testing resourcefulness and creativity. *Advances in Cognitive Systems*, 6:31–44, 2018.

[SS18c] Vasanth Sarathy and Matthias Scheutz. Macgyver problems: Ai challenges for testing resourcefulness and creativity. *Advances in Cognitive Systems*, 6, 2018.

[SS19a] Vasanth Sarathy and Matthias Scheutz. On resolving ambiguous anaphoric expressions in imperative discourse. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[SS19b] Vasanth Sarathy and Matthias Scheutz. On resolving ambiguous anaphoric expressions in imperative discourse. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[SSA+17] Vasanth Sarathy, Matthias Scheutz, Joseph Austerweil, Yoed Kenett, Mowafak Allaham, and Bertram Malle. Mental representations and

computational modeling of context-specific human norm systems. In *Proceedings of Cognitive Science*, 2017.

[SSBFL08] Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. SCARE: A situated corpus with annotated referring expressions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, Morocco, May 2008.

[SSCO08] Lior Shapira, Ariel Shamir, and Daniel Cohen-Or. Consistent mesh partitioning and skeletonisation using the shape diameter function. *Visual Computer*, 24(4):249–259, 2008.

[SSKA07] Matthias Scheutz, Paul Schermerhor, James Kramer, and David Anderson. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4):411–423, 2007.

[SSM17] Vasanth Sarathy, Matthias Scheutz, and Bertram Malle. Learning behavioral norms in uncertain and changing contexts. In *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2017.

[ST85] Glenn Shafer and Amos Tversky. Languages and designs for probability judgment. *Cognitive Science*, 9(3):309–339, 1985.

[ST95] Yoav Shoham and Moshe Tennenholtz. On Social Laws for Artificial Agent Societies: Off-Line Design. *Artificial Intelligence*, 73(1-2):231–252, 1995.

[Sta02] Robert Stalnaker. Common ground. *Linguistics and philosophy*, 25(5-6):701–721, 2002.

[Ste02] Mark Steedman. Formalizing Affordance. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, pages 834–839, 2002.

[Sto03] Thomas A Stoffregen. Affordances as Properties of the Animal-Environment System. *Ecological Psychology*, 15(2):115–134, 2003.

[STS95] S. Shibata, K. Tanaka, and A. Shimizu. Experimental analysis of handing over. *Robot and Human Communication, 1995. RO-MAN'95 TOKYO, Proceedings., 4th IEEE International Workshop on*, pages 53–58, 1995.

[SVAB15a] Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. Towards addressing the winograd schema challenge—building and using a semantic parser and a knowledge hunting module. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[SVAB15b] Arpit Sharma, Nguyen Ha Vo, Somak Aditya, and Chitta Baral. Towards addressing the Winograd Schema Challenge-building and using a semantic parser and a knowledge hunting module. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1319–1325, 2015.

[SW63] Claude E Shannon and Warren Weaver. The mathematical theory of communication. 1949. *Urbana, IL: University of Illinois Press*, 1963.

[SW13] Anna Steidle and Lioba Werth. Freedom from constraints: Darkness and dim illumination promote creativity. *Journal of Environmental Psychology*, 35:67–80, 2013.

[SWK+19] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*, pages 165–193. Springer, 2019.

[SYC+19] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge dataset and models for dialogue-based

reading comprehension. *Transactions of the Association for Computational Linguistics*, 2019.

[Tea70] Mission Evaluation Team. Mission operations report apollo 13. 1970.

[TETC19] Paul Trichelair, Ali Emami, Adam Trischler, and Jackie Chi Kit Cheung. How reasonable are common-sense reasoning tasks: A case-study on the Winograd Schema Challenge and SWAG. *arXiv preprint arXiv:1811.01778*, 2019.

[TFK13] Cristen Torrey, Susan Fussell, and Sara Kiesler. How a robot should give advice. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 275–282. IEEE Press, 2013.

[TGKGM20] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

[THC+16] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends® in Robotics*, 4(2-3):105–223, 2016.

[THPS12] Yuqing Tang, Chung Wei Hang, Simon Parsons, and Munindar Singh. Towards argumentation with symbolic dempster-shafer evidence. *Frontiers in Artificial Intelligence and Applications*, 245(1):462–469, 2012.

[TL09] Laura E Thomas and Alejandro Lleras. Swinging into thought: Directed movement guides insight in problem solving. *Psychonomic bulletin & review*, 16(4):719–723, 2009.

[TM14] Jan Tünnermann and Bärbel Mertsching. Saliency and Affordance in Artificial Visual Attention. *Proceedings of First Workshop on Affor-*

dances: Affordances in Vision for Cognitive Robotics, Robotics Science and Systems, 2014.

[TMS17] Preston P Thakral, Kevin P Madore, and Daniel L Schacter. A role for the left angular gyrus in episodic simulation and memory. *Journal of Neuroscience*, 37(34):8142–8149, 2017.

[TP14] Andreas Ten Pas and Robert Platt. Localizing Handle-Like Grasp Affordances in 3-D Points Clouds Using Taubin Quadric Fitting. In *International Symposium on Experimental Robotics (ISER)*, 2014.

[TPS+20] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374, 2020.

[Tur50] Alan M Turing. Computing machine and intelligence. *MIND*, 59:433–460, 1950.

[Tur92] M T Turvey. Affordances and Prospective Control: An Outline of the Ontology. *Ecological Psychology*, 4(3):173–187, 1992.

[UCS+11] E. Ugur, H. Celikkanat, E. Sahin, Y. Nagai, and E. Oztop. Learning to grasp with parental scaffolding. *2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2011)*, 2011.

[UM77] Edna Ullmann-Margalit. *The emergence of norms*. Clarendon library of logic and philosophy. Clarendon Press, Oxford, 1977.

[UNO13] Emre Ugur, Yukie Nagai, and Erhan Oztop. Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills. *Proc. of the RAAD 2013 22nd InternationalWorkshop on Robotics in Alpe-Adria-Danube Region*, (August 2014):1–19, 2013.

[UNSO15] Emre Ugur, Yukie Nagai, Erol Sahin, and Erhan Oztop. Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese. *IEEE Transactions on Autonomous Mental Development*, (2):119–139, 2015.

[USO12] Emre Ugur, Erol Sahin, and Erhan Oztop. Self-discovery of motor primitives and learning grasp affordances. *IEEE International Conference on Intelligent Robots and Systems*, pages 3260–3267, 2012.

[Var88] Francisco J Varela. Structural coupling and the origin of meaning in a simple cellular automation. In *The semiotics of cellular communication in the immune system*, pages 151–161. Springer, 1988.

[Var15] K.M. Varadarajan. Topological mapping for robot navigation using affordance features. *2015 6th International Conference on Automation, Robotics and Applications (ICARA). Proceedings*, 2015.

[VD16] Kees Van Deemter. *Computational models of referring: a study in cognitive science*. MIT Press, 2016.

[VET13] Jan BF Van Erp and Alexander Toet. How to touch humans: Guidelines for social agents and robots that can touch. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 780–785. IEEE, 2013.

[VRR13] Kathleen D Vohs, Joseph P Redden, and Ryan Rahinel. Physical order produces healthy choices, generosity, and conventionality, whereas disorder produces creativity. *Psychological Science*, 24(9):1860–1867, 2013.

[VTR16] Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind: Cognitive science and human experience*. MIT press, 2016.

[VV11] Karthik Mahesh Varadarajan and Markus Vincze. Knowledge representation and inference for grasp affordances. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6962 LNCS:173–182, 2011.

[Wal26] Graham Wallas. *The art of thought*. Jonathan Cape, London, UK, 1926.

[WASS16] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, 2016.

[WBC+15] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[WBOS15a] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going Beyond Literal Command-Based Instructions : Extending Robotic Natural Language Interaction Capabilities. In *AAAI*, pages 1387–1393, 2015.

[WBOS15b] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond literal command-based instructions: extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[WDPAP12] Rob Withagen, Harjo J De Poel, Duarte Araújo, and Gert-Jan Pepping. Affordances can invite behavior: Reconsidering the relationship between affordances and agency. *New Ideas in Psychology*, 30(2):250–258, 2012.

[WDWK07] Michael L Walters, Kerstin Dautenhahn, Sarah N Woods, and Kheng Lee Koay. Robotic etiquette: results from user studies involving a fetch and carry task. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 317–324. ACM, 2007.

[Weg12] Beeman M. Wegbreit E, Suzuki S, Grabowecky M, Kounios J. Visual Attention Modulates Insight Versus Analytic Solving of Verbal Problems. *Journal of Problem Solving*, 144(5):724–732, 2012.

[Wei85] Karl E Weick. Sources of order in underorganized systems: Themes in recent organizational theory. *Organizational theory and inquiry*, pages 106–136, 1985.

[WFEŠ16] Ricarda Wullenkord, Marlena R Fraune, Friederike Eyssel, and Selma Šabanović. Getting in touch: How imagined, actual, and physical contact affect evaluations of robots. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 980–985. IEEE, 2016.

[Wig06] Geraint A Wiggins. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, 19:449–458, 2006.

[Win80] Terry Winograd. What does it mean to understand language? *Cognitive science*, 4(3):209–241, 1980.

[WK06] Sabrina Wilske and Geert-Jan Kruijff. Service robots dealing with indirect speech acts. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4698–4703. IEEE, 2006.

[WKPW09] K Wickramaratna, M Kubat, Kamal Premaratne, and T Wickramarathne. Rule mining and missing value prediction in the presence

of data ambiguities. In *The Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 2009.

[WS16] Tom Williams and Matthias Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[WS18] Tom Williams and Matthias Scheutz. Reference in robotics: a givenness hierarchy theoretic approach, 2018.

[WSW20] Ruchen Wen, Mohammed Aun Siddiqui, and Tom Williams. Dempster-shafer theoretic learning of indirect speech act comprehension norms. In *AAAI*, pages 10410–10417, 2020.

[WTNS18] Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. Thank you for sharing that interesting fact!: Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 298–306. ACM, 2018.

[XT19] B. Scassellati X. Tan, J. Brawer. That's mine! learning ownership relations and norms for robots. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[Yag87] Ronald R. Yager. On the dempster-shafer framework and new combination rules. *Information Sciences*, 41(2):93–137, 1987.

[Yar67] Alfred L. Yarbus. *Eye movements and vision*, volume 6. 1967.

[YDL+16] Wenjing Yang, Arne Dietrich, Peiduo Liu, Dan Ming, Yule Jin, Howard C Nusbaum, Jiang Qiu, and Qinglin Zhang. Prototypes are Key Heuristic Information in Insight Problem Solving. *Creativity Research Journal*, 28(1):67–77, 2016.

[YLS+16] Chao Yu, Hongtao Lv, Sandip Sen, Jianye Hao, Fenghui Ren, and Rui Liu. An Adaptive Learning Framework for Efficient Emergence of Social Norms (Extended Abstract). In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1307–1308, 2016.

[YR14] Sureyya Yoruk and Mark A Runco. Neuroscience of Divergent Thinking. 56(1):1–16, 2014.

[Zad79] Lotfi A Zadeh. *On the Validity of Dempster's Rule of Combination of Evidence*. Electronics Research Laboratory, University of California, 1979.

[ZBSC18] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

[ZHL+17] Philipp Zech, Simon Haller, Safoura Rezapour Lakani, Barry Ridge, Emre Ugur, and Justus Piater. Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior*, (5):235–271, 2017.