RI: Small: Reasoning about Containers: Cognitive and Automated Models

Ernest Davis and Gary Marcus, New York University

Project Summary

Computers surpass humans in many respects, but remain notoriously poor at common sense reasoning, such as reasoning about every day physical objects. In computers this is generally carried out through the use of *step-by-step physical simulation*; a computer program is given an exact specification of a physical situation, and the computer calculates the precise trajectory of the system at a sequence of discrete time points. There is good reason, however, to think that human beings may use different techniques -- and that the techniques of humans may offer significant advantages over simulation-based calculation. First, humans can make useful qualitative predictions given only partial specifications of the physical and geometrical properties of the situation. Second, humans can make useful qualitative predictions even if they have only a very imperfect knowledge of the physics of the objects or materials involved. Third, humans can predict the qualitative behavior of a complex situation without needing to calculate all the details of the behavior. Finally, humans can use the same physical knowledge for a wide variety of cognitive tasks, including not just prediction but also manipulation, planning design, vision, and text understanding. Our goal is to develop a theory that explains how reasoning with these characteristics can be carried out, both in humans and machines.

We aim to study qualitative physical reasoning of this kind from two directions: psychology and artificial intelligence (AI). First, we will carry out experimental studies of human commonsense physical reasoning, and we will develop cognitive models based on the results of those studies. Second, we will develop an AI system that will use symbolic reasoning to carry out qualitative physical reasoning. The two approaches will be synergistic: the results of the experimental study will guide the construction of the AI system; conversely, the AI system will serve, both as a proof of concept and as a source of insights and constraints for the cognitive model.

Our particular empirical focus will be one of the fundamental challenges in physical reasoning: understanding the relations between containers such as boxes, cups, bags, cages, automobiles, and so forth, and their contents. Such relationships are ubiquitous in everyday life very well understood by human reasoners, and span a very wide range of shapes and materials. The contents of a container can be essentially any shape and any material; the container itself may have a wide range of shapes and materials. (Standard physical simulations, in contrast, are typically restricted to certain classes of materials, effective with solid-bodies, but also less effective with certain kinds of malleable materials.)

Intellectual impact: By synergistically combining psychological and computational studies, this project has the potential to both contribute substantially to our understanding of commonsense physical reasoning as a cognitive process, and to the state of the art in automated commonsense physical reasoning.

Broader impact: In addition to training graduate students, we aim to significantly enhance the general public's understanding of the challenges of building artificial intelligence with common-sense reasoning, through writings in the popular media (Marcus is currently blogging for *The New Yorker*, and has also written for *Wired,* the *Wall St Journal*, and the *New York Times*), and, we hope, through an exhibition on the relation of cognitive psychology and artificial intelligence at a major science museum (see letter of support from Paul Hoffman).

Keywords: Physical reasoning, commonsense reasoning, cognitive models, simulation, qualitative reasoning.

RI: Small: Reasoning about Containers: Cognitive and Automated Models

Proposal Submitted to the NSF, "Robust Intelligence"

Ernest Davis, Department of Computer Science, New York University, New York, NY

Gary Marcus, Department of Psychology, New York University, New York, NY

1. Qualitative physical reasoning

Computers surpass humans in many respects, but remain notoriously poor at common sense reasoning, such as reasoning about every day physical objects, Predicting the behavior of physical systems has , of course, been one of the central objectives of mathematics and computer science since the inceptions of each field to the present day; and techniques for carrying out such predictions comprise a significant fraction of the content of both fields. Programs exist that can create extremely detailed simulations of the interactions of 200,000,000 deformable red blood cells in plasma (Rahimian et al. 2010); the air flow around the blades of a helicopter (Murman et al., 2003); the interaction of colliding galaxies (Benger 2008); and the injuries caused by the explosion of an IED under a tank (Tabiei and Nilakantan). Software, such as NVidia PhysX, that can simulate the interactions of a range of materials, including rigid solid objects, cloth, and liquids, in real time, is available for the use of game designers as off-the-shelf freeware (Kaufmann and Meyer, 2008).

Still, most of the mathematics and nearly all of the computer science applies only in conditions where there is complete information available. Exact physical equations for the physical interactions must be known, and the problem specifications must give precise values for the material characteristics, physical properties, and geometric relations of all the objects and materials involved. (We will discuss exceptions in section 5.2 below.)

Human reasoners, in contrast, often make sound inferences about the world, rapidly, with far less complete information. Consider, for example, the following scenario: You take a half-filled coffee cup and turn it upside down. Given exact specifications of all the elements involved — the shape of the cup, the amount of liquid in the cup, the exact motion involved in turning the cup, the distance from the floor — a simulation program can produce a precise[1] prediction for what will happen. A human reasoner, by contrast, can predict that the result will be a puddle of coffee on the floor, underneath the cup, two or three feet in diameter; and she can predict this with only very approximate knowledge of the shape of the cup and so on. On the other hand, the reasoner is not simply using a specific rule of the form "Turn coffee cup upside down ⇒ puddle on floor", as is evidenced by the fact that she can take into account all kinds of circumstances that modify the conclusion. If the cup is over the center of a table, then the puddle may end up entirely on the table; if the cup is over the edge of a table, the puddle will probably end up partly on the floor and partly on the table. If the floor is slanted, then the coffee will continue to run in the direction of the slant. If the cup is covered, then the coffee will not pour; if the cover has a small hole, then it will leak slowly out the hole. If there is a fan blowing hard toward the stream of coffee, then the coffee will be deflected.

[1] The prediction will be *precise,* but how *accurate* it will be is a different and much more difficult question.

No existing program can carry out this kind of reasoning. The goal of this grant is to glean insight into how humans perform rapid approximate physical reasoning with incomplete information, and to use insights from human intuitive physics to help develop better systems for automated physical reasoning.

In particular, we will focus on four characteristics of human commonsense reasoning:

- The ability to reason with incomplete spatial and physical specifications.
- The ability to reason about different physical materials whose characteristics or interactions are imperfectly understood.
- The ability to reason easily and quickly about broad qualitative features of situations whose detailed behavior is complex.
- The ability to use a piece of knowledge for many different kinds of cognitive tasks, such as prediction, planning, explanation, and natural language understanding,

2. Containers and Contents

The test bed that we will use for studying these characteristics of commonsense physical reasoning will be the relations between containers — boxes, bags, cups; wombs, cribs, rooms, and so forth — and their contents. Such relations are ubiquitous in everyday life, and very well understood by people; this understanding is fundamental to spatial and physical reasoning of all kinds.

As an important step towards improving machine understanding of common sense, we propose to study how people understand, use, and learn the meaning of these relationships and how an understanding of these cognitive underpinnings can lead to more robust methods for automated commonsense reasoning.

A container can serve many different purposes:

- To carry contents that are difficult or impossible to carry directly. For example, a shopping bag or a bottle.
- To ensure that the contents remain in a fixed place. For example, a crib or a cage; a cup that remains stationary.
- To protect the contents against other objects or physical influences. For example, a case or a thermos bottle.
- To hide the contents from inspection. For example, an envelope.
- To ensure that objects can only enter or exit through specific portals. For example, a tea-kettle.

In some cases it is necessary that some kinds of material or physical effects can either fit through the portals or pass through the material of the container, while others cannot. For instance, a pet carrying case has holes to allow air to go in and out; a display case allows light to go in and out but not dust.

There are four primary kinds of physical principles involved in all of these cases. First, matter must move continuously; if the contents could be teleported out of the container, as in *Star Trek*, none of these would work. Second, the contents (or the externality being kept out, such as dust) cannot pass through the material of the container. Third, there are constraints on the deformations possible to the shapes of the container and of the content. Fourth, in the case of an upright open container, gravity prevents the contents from escaping.

Simple, natural examples of commonsense physical reasoning reveal a number of important characteristics.

First, human reasoners can use very partial spatial information. For example, consider the text, "There was a beetle crawling on the inside of the cup. Wendy trapped it by putting her hand over the top of the cup, then carried the cup outside, and dumped the beetle out onto the lawn." A reader understands that the cup and the hand formed a closed container for the beetle, and that Wendy removed her hand from the top of the cup before dumping the beetle. Thus, the reader uses general spatial knowledge about cups, hands, and beetles in interpreting the text and does not require the geometry of these to be specified precisely. Physical simulators can calculate the interactions of all these objects only when their shapes are precisely specified.

Second, human reasoners can in many cases infer that a material is confined within a closed container even if they have only a vague idea of the physics of the material of the container and almost no idea at all of the material of the contents. By contrast, automated reasoning systems that rely on detailed physical simulations are typically far less robust. Simulation systems such as NVidia PhysX can deal effectively and efficiently with solid objects and other specific materials such as liquids or cloth. However, materials in these systems are understood either completely or not at all. If a material is outside its repertoire, it cannot fail gracefully, carrying out partial reasoning; it fails to give an answer at all. For this reason, a model of reasoning about containers that relies on having detailed physical models of the materials of the container and of the contents is entirely implausible as a cognitive model and is inadequately general as a model for automated reasoning.

Third, human reasoners can predict qualitative behavior of a system and ignore the irrelevant complex details; unlike much software, they are often very good at seeing the forest and not being distracted by the trees. For example, if you pour water into a cup, you can predict that, within a few seconds it will be sitting quietly at the bottom of the cup; and you do not need to trace through the complex trajectory that the water goes through in getting to that equilibrium state.

Finally, knowledge about containers, like most high-level knowledge, can be used for a wide variety of tasks in a number of different modalities, including prediction, planning, manipulation, design, textual or visual interpretation, and explanation. The container relation is also often used metaphorically; e.g. for the relation between a memory location and a value in computer science.

3. Proposed Research

Our proposed research project will study both the empirical psychology of adult humans and the design of an automated intelligent system.

1. **Cognition:** How do humans reason about the relationships between containers and their contents? In particular, how do they carry out reasoning in cases where only partial information is available, and how are the reasoning methods used in these general cases integrated with the more specialized techniques available where more constraints are known? To what extent does reasoning rely on broad general principles and to what extent on special-case rules? To what extent do different cognitive tasks draw on the same general knowledge versus employing task-specific heuristics? What sorts of problems if any lead human reasoners to erroneous inferences about containers?
2. **Automated reasoning:** How can an automated reasoner be constructed that achieves the same kind of flexibility in reasoning shown by humans?

3.1 Adult psychology

In experimental work, we plan to investigate how human adults solve certain physical reasoning problems, contrasting inferences that might be derive from rule-based heuristics from inferences drawn from direct physics-engine like simulations. Our primary focus will be on four features that distinguish these two categories of models.

Step-by-step simulation vs. heuristic characterization over extended time. A simulation model necessarily computes the entire trajectory of the physical system involved from start to finish. A rule-based system can take advantage of heuristics that characterize the end state of a system, or partially characterize the trajectory, without computing the intermediate states in detail. For example, an experimenter picks up a closed bottle of water, and say to the subject, "I am going to shake the bottle and then put it down on the table. What will happen to the water?" A subject should be able to predict that the water will remain in the bottle while it being shaken, and then, once the bottle is at rest, will go back to sitting at the bottom of the bottle with a horizontal upper surface. In a rule-based system, this prediction requires tracking the complicated motion of the water from the shaken state to the rest state.

Inference from precise problem specifications vs. partial information. A simulation model is fundamentally based on reasoning with precise fully detailed information. Reasoning with partial information can be performed only by doing a Monte Carlo search over the space of precise instantiations of the partial information. Rule-based systems are inherently designed to deal with partial information; precise information is just a special case of partial information, and not necessarily a particularly tractable special case. For example, in the same experiment with shaking the bottle, in a simulation model, the reasoner must simulate a large number of different possible motions that might be involved in the shaking motion. If the shape of the inside of the bottle is unknown, then the reasoner must simulate the situation with a variety of different possible inner shapes. The direct effect of this need for multiple simulations on subjects' response times is hard to predict exactly, since alternative simulations can in principle be carried out in parallel. However, one way or another, the need to reason with partial information must increase the demand on cognitive resources of some form. The more incomplete the information and the more forms of incomplete information are involved (e.g. position, shape, mass, motion, material, number of objects), the larger this increase should be.

Full vs. partial knowledge of materials. A simulation model requires a fully detailed model of all the materials involved; without that, no predictions at all can be made. A particularly vivid example of this is in dealing with animals. A reasoner presumably does not have a useful mechanical model of the animal; nonetheless she can do some kinds of reasoning about it. For instance, a reasoner can predict that an eel in a fish tank will remain in the tank, even without having any idea what the mechanics are of the eel's locomotion, whereas a simulation would, in the absence of a detailed model of an eel, be unable to make any prediction at all. (Davis and Marcus, in prep.).

Collections. In a physical system with many similar objects, such as a pail full of sand, or with many similar shape features, such as a sieve with many holes, a system based on simulation must, first, do a Monte Carlo search with random instantiations of the number of such objects and features, and, second, in each such instantiation, reason about each grain of sand or each hole through the sieve individually. In a rule-based system, it is at least in principle possible to reason collectively about the grains of sand or the holes through the sieve (though admittedly this is a challenging problem for automated reasoners). In a rule-based system, if sets of related objects are viewed as collections rather than individuals (Halberda, Sires, and Feigenson, 2006) increasing the number of objects should have little effect; e.g. there should be essentially no difference between reasoning about 100 grains of sand and reasoning about 1000. In a simulation-based system, the difficulty of reasoning continues to increase: reasoning about 1000 should be more difficult than reasoning about 100 in the same way that reasoning about 10 is more difficult than reasoning about 1.

To study how human reasoning deal with challenges of these sorts, we will present subjects with 48 scenarios such as those described below, drawn from four categories, presented in random order, counterbalanced across subjects, as videos that will be rendered in advance using a physics engine. Subjects will be requested to respond as quickly as they can while still being accurate. We will measure their response time and analyze their answers in terms of speed and accuracy.

A sample scenario for each of the four features:

Scenario 1: Full trajectory vs. prediction over extended time
Subjects are shown a small die being dropped into a funnel with steep sides held over a table (Davis, 1988). They are asked where the die will be when it comes to rest. There are two different sizes of funnels; the top, conical, section of the larger funnel is the same shape as the smaller funnel but larger in each dimension; the bottom, cylindrical section is the same diameter in the two funnels, but longer in the larger funnel. In each case the original position of the die is just below the top of the funnel, well off-center.

A rule-based theory predicts that both situations will have the same response time, since the qualitative information is the same.
A simulation-based theory will predict that it will take longer to simulate the situation in the larger funnel; hence, the response time should be longer. This difference should be manifested, whether the simulation process uses a constant time increment (e.g. the state of the die is calculated at 100 millisecond intervals), since the actual time required to fall through will be greater in the second case, or it uses a variable time increment, tracking from one significant event to the next event, since the fall of the die through the larger funnel will both take more total time and involve more collisions and changes of contact.

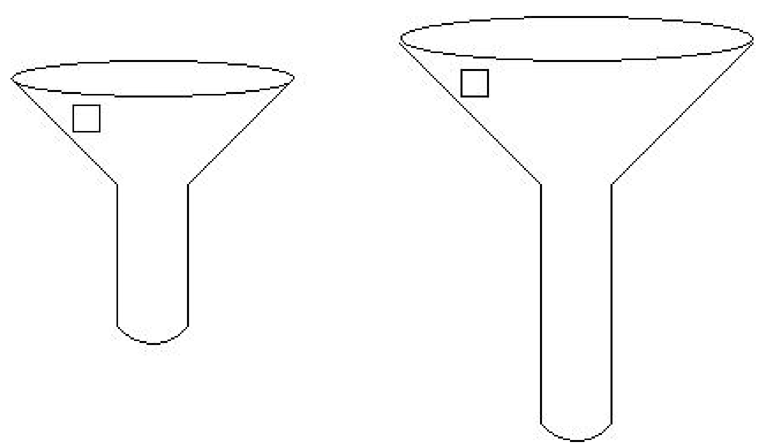

Figure 1: Scenario with a die being dropped through a funnel

An alternative version of this example will show subjects the two situations, placed over a table such that the starting position of the die is at equal heights off the table, and ask them to predict in which situation it will take *longer* for the die to reach the table (as opposed to simply inquiring about the final state). A correct response -- that it will take longer for the longer funnel, -- would suggest that subjects are using a physics-like simulation, rather than rules, inasmuch as such an inference is easy to derive from simulation and extremely difficult to get from a rule-based system (such as Davis, 1988).)

Scenario 2: Partial vs. complete Information
Subjects are shown one of three situations:

> Situation 1: There are 3 boxes: A, B1, and B2, each with a lid. The experimenter shows that A will fit into B1, but not into B2. Subjects are told, "There is an object O in A. If I take O out of A. which of the other boxes can I be sure it will fit into?"
> Situation 2: Boxes A, B1, B2 are the same as in situation 1. After showing that A fits in B1, but not in B2, the experimenter shows object O inside A. Subjects are asked which of the other boxes O will fit into. The shape of O is quite complicated.
> Situation 3: There is no box A. Object O is shown outside boxes B1 and B2. Subjects are asked which one it fits into. The shape of O is the same as in situation 2, so that in fact O fits inside B1 and B2,

We can divide the reasoning tasks into three categories:

i. In situation 1, since the subjects do not see O, they cannot use precise shape information; they must make the inference that, since O fits in A and A fits in B1, O fits in B1.
ii. In situation 3, since there is no easy qualitative inference like the one in situation 1, subjects must use the precise shapes of O and the boxes.
iii. In situation 2, to infer that O fits in B1, subjects may either use the exact shapes of O and B1 or may use the inference used in situation (1). To infer reliably that O does not fit in B2, they must use the exact shapes of O and B2.

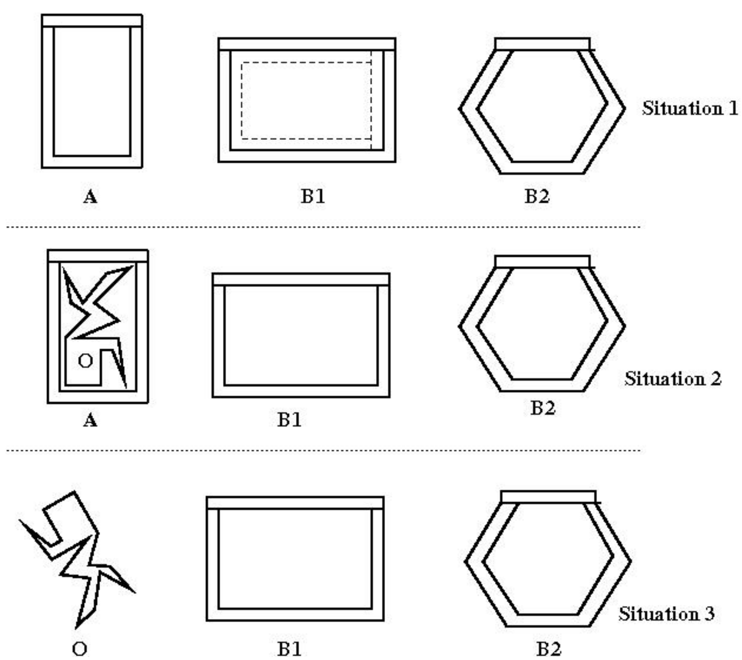

Figure 2: Scenario 2

If subjects are using a simulation model in reasoning then "necessarily use precise" should be faster and more accurate than "necessarily use partial"; if they are using rule-based reasoning, then "necessarily use partial" should be faster and more accurate than "necessarily use precise". Moreover, under the latter assumption, if the behavior for the "either method" problems is more similar to the "necessarily use precise" problems, then one can infer that subjects preferentially use precise information when available;

if the behavior is more similar to the "necessarily use partial" problems, one can infer that they preferentially use partial information. (Figure 2 shows a two-dimensional version; the actual experiment will involve three-dimensional objects, so the precise judgments will be significantly more difficult than figure 2 suggests.) Rule-based reasoning that can support this kind of inferences is discussed in (Christani 1999) and in (Davis 2013).

Scenario 3: Partially known dynamics

Subjects are shown one of six situations:

Situation 1: There is a small ball in a closed, sealed box.

Situation 2: There is a slinky inside the box.

Situation 3: There is a grasshopper inside the box .

Situations 4-6. Identical, except that the box is open on top.

Subjects are asked, "If the box is shaken hard up and down, can the ball/slinky/grasshopper come out of the box?"

If the subjects are using a rule-based system, then the three problems should be of comparable difficulty. If the subjects are using a simulation-based system, then prediction should be much more difficult in situations 2,3,5, and 6, because the dynamic model of a slinky or a grasshopper are much more complicated, and much less well known to the subject, than the dynamics of a ball. Subject's predictions should therefore be slower and less reliable. (The comparison between slinky and grasshopper allows us to begin investigate the role of animacy in physical reasoning.)

Scenario 4; Collections of objects

Subjects are shown one of three situations:

Situation 1: There is a single marble in a bottle, much smaller than the opening of the bottle

Situation 2: There are 10 such marbles in the bottle

Situation 3: There are 50 such marbles in the bottle,

Subjects are asked: If I turn the bottle upside down, will the marble(s) fall out?

If the subjects are using a rule-based system, then situation 2 will be only somewhat more difficult than situation 1, and situation 3 will be only negligibly more difficult than situation 2.

If the subjects are using a simulation-based system, then situation 2 will be much more difficult than situation 1 and situation 3 will be much more difficult than situation 2, because the simulation must track the interactions of all the individual marbles.

Further scenarios will address other issues, such as the nature of reasoning about different materials,

3.2 Automated reasoning

For the automated reasoning portion of this project, we will develop a system that can reason with a range of more and less specific physical and geometric properties and draw reasonable inferences. That system will consist of a representational language that can express partial information about physical situations and a rule-based system that can carry out various kinds of reasoning tasks based on the representation. The emphasis in this part of the project will be on dealing adequately with weak information, as humans do.

Specifically, we plan to develop a representation language that can deal with a wide range of materials, to describe under what circumstances they can serve as a container, and under what circumstances they can either pass through the material of a container or fit through the holes of a container. Materials include rigid objects; paper; cloth; string; liquids; gasses; light (as a thing to be blocked); animals (as contents); and human hands (as containers when cupped).

The representation language will need to express spatio-temporal information, physical information, and information about actions. Our representation language will be in the spirit of the work that has been done in qualitative reasoning (e.g. Bobrow 1985), and qualitative spatial reasoning (e.g. Cohn and Renz 2008), discussed below in section 5.2. However in our project we will address additional issues that arise in reasoning about the container/content relation, such as whether one object fits through a portal or whether a malleable object will fit in a space, or whether a malleable object can be cupped around a cavity, that previously have received little or no attention in the work on qualitative spatial reasoning. Working out a complete specification of the representation language is one aspect of the proposed project, but broadly speaking, we are planning on using a constraint language with the following elements:

1. Spatial language
    a. Categories of entities: Extended regions, measures of distance, height, and volume.
    b. Mereological relations over regions: "A is part of B", "A overlaps B", "A and B are disjoint".
    c. Contact relations, such as "The boundaries of regions A and B meet."
    d. Euclidean congruence, for rigid objects: "Regions A and B are congruent".
    e. Topology of holes: "A is an inner cavity of B," "H is a hole through A connecting inner cavity C to the outside".
    f. Comparative distance, height, or volume: "X is greater than Y".
    g. Characteristic measures associated with regions; "D is the diameter of region R"; "D is the inner radius of region R", "V is the volume of region R".
    h. Heights associated with regions: "H is the height of the bottom/top of region R."
    i. Cylinders: "C is a right circular cylinder of radius R and length L"
2. Temporal language
    a. Categories of entities: Instants of time and fluents.
    b. Order of instants: "Time I is earlier than J"
    c. Relation of instants to fluents: "Fluent F holds at time I"; "F holds throughout interval [I,J]".
    d. Preconditions on actions: "Action A is possible at time I if conditions Q1, Q2 ... hold."
    e. Effects of actions: "If action A is carried out then Q will become true."
3. Spatio-temporal language:
    a. Category of entities: histories (= region-valued fluents).
    b. Constraints: "Spatial relation Q holds on history A at time I / over time interval [I,J]".
    c. Passing through: "History A passes through.history B during time interval [I,J]"
4. Physical language
    a. Categories of entities: Objects, materials
    b. Relations between histories and physical entities: "History H corresponds to the position of object O", "History H is filled with material M", "History H is empty".
5. Actions. For simplicity, we will abstract away the actual manipulator used for actions, and model actions as if the agent could move objects by telekinesis. Under that assumption, an action amounts to the trajectory (history) followed by the manipulated objects.
    a. Category of entity: Histories
    b. Occurrence relation: "Action A occurs over time interval [I,J]".

Problems we leave for future research include integrating this qualitative language with a language of precise geometric and material specifications; characterizing complex shapes such as the range of shapes that a piece of paper can be crumpled into or that a human hand can attain; and dealing with collections of objects and shape features. The problem of complete, general inference over a language as expressive as sketched above is of course hugely intractable (almost certainly undecidable). However, we do not need, and certainly do not intend, to implement a general inference engine, just an inference engine that carries out the particular inferences needed for commonsense reasoning about containers.

The reasoning system will be able to reason about the interactions between objects described in the language of materials and shapes. For example, it will:

- Predict that the contents of a closed container will remain inside the container.
- Infer that, if an object fits inside a cavity formed by a box and a lid, then it is possible to place the object in the box and then close the lid.
- Predict that a rigid object will not fit in a cavity if the diameter of the object is greater than the diameter of the cavity.
- Predict that fluids can pass through a portal of any size or shape.
- Predict that, for any object, a sufficiently large box will contain the object.

Using this representation language, we will assemble a knowledge base sufficient to support most or all of the inferences based on partial information that we will study in our experimental scenarios, except for those dealing with collections, which we leave for later study.

In the initial implementation of our automated reasoner, we plan to use a first-order theorem prover such as SPASS (Weidenbach et al. 2002) or Prover9 (McCune). This technology has been applied very effectively in a wide range of applications ranging from qualitative spatial reasoning (Wölfl, Mossakowski, and Schröder 2007) to program verification (e.g. Cook, Kroening, and Sharygina 2004). However, as the project progresses and we learn more about the kinds of reasoning involved in these cognitive tasks, we may either supplement the theorem prover with such features as support for default reasoning (Brewka, Nielmelä, and Truszczński 2008)) or higher-level control heuristics (Cox and Raja 2011); or adopt a different reasoning architecture such as answer-set programming (Gelfond 2008).

A special focus, both computationally and empirically, will be on the likelihood that human reasoners may use collections of mutually inconsistent partial theories in dealing with problems of this kind, and an automated reasoner that is intended to serve as a cognitive model must to some extent address the issues of dealing with inconsistent beliefs. How this can best be done is a matter for research, and certainly substantially depends on the findings of our psychological studies. Our initial plan is to use a truth-maintenance system (Forbus and de Kleer, 1993); again, these have been applied successfully both in qualitative reasoning (Forbus and de Kleer 1993) and in program verification (where the technique is known as clause learning; Gomes et al. 2008).

4. Evaluation

The AI segment of the project will be evaluated according to three kinds of criteria:

- Representational adequacy. To what extent can the representation express the kinds of partial knowledge that arise in commonsense reasoning about containers?
- Inferential adequacy. To what extent can commonsense inferences about containers be carried out in this system?
- Cognitive modeling. To what extent can features of human reasoning be characterized in terms of this model? As remarked above, this will be inherently limited by the fact that we do not intend to address reasoning from inconsistent knowledge in our AI system.

For representational adequacy and inferential adequacy, since there does not exist either a standard collection of benchmarks for this domain or, as far as we know, a ecologically natural source of problems, An important objective will be to develop a suite of benchmark problems, along the lines of the scenarios described in the experiments above, probably numbering in the hundreds, that will be empirically tested against human performance, and be freely distributed to serve as a testbed for competing approaches to physical reasoning.

5. Related work

Our project builds on existing bodies of work in cognitive psychology and in artificial intelligence.

5.1 Related work in cognitive psychology

Several leading researchers in developmental psychology, including Renée Baillargeon, Susan Hespos, and Elizabeth Spelke (Hespos & Baillargeon 2001; Hespos & Spelke 2007), have examined young infants' understanding of containers, establishing that within the first six months of life, infants have some basic understanding of the fact that objects can be hidden within containers . Further work has shown that by the end of the first year of life, infants develop a more refined concept, recognizing for example that it is surprising that an object A (say a small hat) conceals an object B that is bigger than A (say, a large rabbit). Less work has been done in adults understanding of containment, and we are not aware of any studies in adults that aims to do what we aim to do here, viz characterizing the specific cognitive processes underlying human common-sense reasoning in situations relating to containers and their contents,

An older line of research due to Piaget, well-known but now somewhat controversial, concerned children's understanding of the “conservation” of liquid, matter, number and so forth. For example, in classic tasks, six years seemed confused about what happened when the liquid in a tall skinny container was poured into a short wide container (Inhelder & Piaget, 1959); later work raised questions about whether children's difficulties stemmed a genuine lack of physical understanding (e.g. Siegal, 1991).

Téglás et al. (2007, 2011) have studied infants' expectations about the time required for objects to escape from a container with holes, and have shown that the level of surprise that the infants exhibit, as indicated by staring time, corresponds to the predictions of a simulation-based model.

Hamrick, Battaglia, and Tenenbaum (2011) studied adult physical reasoning in predicting stability of a tower of blocks, and showed that the data matched a simulation-based model that incorporates a probabilistic element corresponding to the uncertainty of the reasoner's perception. Marcus and Davis (submitted) show that the models make incorrect predictions, however, in another closely-related task, and earlier work by McCloskey (e.g., 1983) and Hecht & Proffitt (1995) shows that people's intuitive physics is not always veridical.

Markman, Klein and Suhr (2009) survey the use of simulation-based models in psychology.

5.2 Related work in artificial intelligence

The best-known body of work on physical reasoning with partial information are the techniques pioneered in the seminal programs QSIM (Kuipers 1985), QP (Forbus 1985) and ENVISION (de Kleer and Brown

1985);[2] since these early papers, these techniques have been very much extended and have been applied to a wide range of physical systems (Forbus 2011). However, this approach is for the most part limited to physical systems whose state can be characterized in terms of a collection of one-dimensional parameters and inequalities between those parameters and landmark values in the parameter space; and whose dynamics can be characterized in terms of a set of qualitative differential equations. In particular, these techniques do not apply well to the kind of reasoning about spatial relationships and spatial change that are central in reasoning about containers.

More relevant to our project is the substantial literature on qualitative spatial reasoning, initiated by the papers of Randall, Cui, and Cohn (1989) and of Egenhofer and Franzosa (1991) and extensively developed since (Cohn and Renz, 2008). These do indeed provide a language that can express some of the qualitative relations needed for our theory and a set of rules that can justify some of the inferences we wish to carry out; and we will certainly be building on these. For instance, in a language that combines the RCC-8 (Region Connection Calculus) calculus (Randell, Cui, and Cohn 1989) with set operations, the relation "*A* is contained inside container *C*" can be expressed as the quantifier-free formula $P(A,B) \wedge NTPP(B,D) \wedge C = D \setminus B$; here *B* is the entire interior cavity, *D* is the union of the container with the cavity (figure 1). $P(A,B)$ is the relation "*A* is a subset of *B*"; $NTPP(B,D)$ is the relation "*B* is a subset of the interior of *D*" and $D \setminus B$ is the normalized set difference *D* minus *B*. Likewise, the relation, "Rigid solid object *A* fits inside interior cavity *B* of closed container *C*" is expressed in the formula $CGPP(A,B) \wedge NTPP(B,D) \wedge C = D \setminus B$, where $CGPP(A,B)$ is the relation "*A* is congruent to a partial part of *B*", defined in (Christiani 1999). However, we will need to extend this theory substantially, as discussed in section 3.3 above.

Davis (2008, 2011). has developed representation languages and systems of rules to characterize reasoning about loading solid objects into boxes and carrying objects in boxes, and pouring liquids between open containers. This work is obviously closely related to the proposed project, and we will draw on it extensively. However, the proposed project differs in significant respects from these earlier works:

- The new project deals with a wider range of materials. (For reasons of computational efficiency, we will not include a rich theory of dynamics of solid and liquid motion — for instance, the analysis of pouring liquid by tilting a cup.)
- The new project aims toward a theory that is both effectively implementable and cognitively realistic, sacrificing expressive and inferential power where necessary.

6. Results from Prior NSF Support

Grant: "Automating Commonsense Reasoning for Elementary Physical Science," NSF IIS-0534809, $328,877, 2/06-8/10. PI: Ernest Davis

In research supported by the above NSF grant, Davis and his associates carried out in-depth studies in physical reasoning and qualitative spatial reasoning His research group also carried out research developing a number of techniques for improved retrieval of web documents.

Another substantial educational project supported by the grant was the development of a new course, "Mathematical Techniques for Computer Science Applications", an introductory course in linear algebra,

[2] Regrettably, the term "qualitative reasoning" is often applied narrowly to theories of this specific form, which leads to terminological difficulties when contrasting this particular class of techniques with other methods of qualitative reasoning

probability, and statistics for computer science masters students, and the writing of a textbook for the course (Davis, 2012).

Physical Reasoning:

Our studies of physical reasoning have led to the following results,

1. The analysis of commonsense reasoning about loading objects into boxes and carrying objects in boxes. (Davis, 2011)

2. The analysis of commonsense reasoning about carrying liquids in containers and pouring liquids between containers. (Davis, 2008)

3. The formulation of a collection of inferences about simple physical and chemical process in a number of alternative ontologies of matter: An ontology based on particles, one based on fields, one based on histories, one based on chunks, one based on infinitesimal particles, and one hybrid ontology combining particles, fields, histories, and chunks. (Davis, 2010, and in prep.)

Spatial Reasoning

4. An analysis of the expressivity of the first-order language allowing quantification over regions, and containing the one predicate, "*Closer(x,y,z)*" (region *x* is closer to *y* than to *z*). We have show that any relation that is analytical and invariant under orthogonal transformations can be expressed in this language. Roughly speaking, the language is capable of expressing essentially all the concepts in standard mathematical geometry and analysis. (Ref) Similarly, the first-order language over the same domain containing the two predicates "*C(x,y)"* (*x* is connected to *y*) and "*Convex(x)*" (*x* is convex) can express any analytical relation that is invariant under affine transformations (Davis, 2006)

5. An analysis of a number of techniques for reconstructing spatial regions from sample points, and a proof that, under specified conditions, the reconstructed region is ``close'' to the true region, under a number of different definitions of ``closeness.'' (Davis, 2012a)

6. An analysis of the use of transition graphs in reasoning about continuous spatial change. We give general definitions of different categories of transition graph for a partition of a topological space. We prove that the class of paths through the graphs is elementary equivalent to the class of continuous paths through the space, relative to a specified first-order language. We show how this theory can be applied in real-world domains such as rigid objects, strings, and liquids. (Davis, 2012b)

Web Search Engines

7. As a doctoral thesis, Ziyang Wang developed and tested a system that monitors a local web site for new information and presents it to the user (Wang, 2006).

Development of human resources

During the period of NSF support, one student completed a doctorate under Davis' advisement::
Ziyang Wang, ``Incremental Web Search: Tracking Changes in the Web.'' May 2006.

Publications

E. Davis. "The Expressivity of Quantifying over Regions." *Journal of Logic and Computation*, **16**, 2006, 891-916.

E. Davis. "Physical Reasoning." In *The Handbook of Knowledge Representation*, F.van Harmelen, V. Lifschitz, and B. Porter (eds.), Elsevier, Oxford, 2008, chap. 14, pp. 597-620.

E. Davis."Pouring Liquids: A Study in Commonsense Physical Reasoning." *Artificial Intelligence*, **172**, 2008, pp. 1540-1578.

E. Davis. "Ontologies and Representations of Matter." *AAAI-10*

E. Davis. "How Does a Box Work? A Study in the Qualitative Dynamics of Solid Objects." Artificial Intelligence. **175,** 2011**,** 299-345.

E. Davis. "Preserving Geometric Properties in Reconstructing Regions from Internal and Nearby Points." *Computational Geometry: Theory and Applications*, **45,** 2012, 234-253

E. Davis, Qualitative Reasoning and Spatio-Temporal Continuity, in S. Hazarika ed. Qualitative Spatio-Temporal Reasoning and Representation: Trends and Future Directions, IGI Global 2012.

E. Davis, *Linear Algebra and Probability for Computer Science Applications*, CRC Press, 2012, 431 pp.

E. Davis, *The Logic of Coal, Iron, Air, and Water: Representing Common Sense and Elementary Science*, in preparation.

Z. Wang, *Incremental Web Search: Tracking Changes in the Web*. NYU Ph.D. thesis, May 2006.

8. Intellectual merit

8.1 Advancement of knowledge.

We expect the project to advance our understanding of commonsense physical reasoning as a cognitive process; to advance the state of the art in automated commonsense physical reasoning; and to serve as an example of how psychological and computational studies of high-level reasoning can be pursued in tandem synergistically. Specifically, we will study reasoning about the interactions of containers and their contents in adults, in children, and in automated reasoners. The reasoning tasks that we will consider will involve much broader classes of materials, partial geometric specifications, and directions of inference than have previously been considered in the related psychological or AI literature.

The deliverables of the project will include:

- Experimental studies of commonsense reasoning about containers in adults,
- Cognitive models that account for the experimental results.
- A representation language for the relevant kinds of partial physical and spatial knowledge.
- A knowledge base that expresses commonsense knowledge of containers and that supports commonsense inferences.

- A reasoning architecture for the knowledge base that can carry out commonsense reasoning in this domain.

The techniques developed and information gained for this rather narrow though important domain will presumably generalize to insights that can be used in theories of high-level cognition and automated reasoning for other kinds of commonsense reasoning as well.

8.2 Plan of work

1. Preliminary stages (prior to support and early parts of Year 1 of support):
   1.1. Construction of a corpus of inferences to be used for evaluation and for guiding theory development.
   1.2. Design of initial experiments
2. Core project (Later parts of year 1, years 2 and 3). A cyclical feedback loop, in which all of the following tracks are pursued in parallel, and results from one track are continually used to guide all the other tracks:
   2.1. Running and evaluating experiments.
   2.2. Development of cognitive models
   2.3. Design of new experiments
   2.4. Development of representation, theories, and inference engine
   2.5. Evaluation of automated system
3. Project conclusion (end of year 3)
   3.1. Overall evaluation of cognitive models, and analysis of outstanding problems.
   3.2. Overall evaluation of automated reasoner, and analysis of outstanding problems.
   3.3. Implications for other forms of high-level cognition and automated reasoning.

8.3 Qualifications of the Principal Investigators

Ernest Davis has been working for almost thirty years in the areas of commonsense physical reasoning and qualitative spatial reasoning for AI systems, and has published extensively in those areas. In particular, he has recently authored two papers bearing directly on qualitative reasoning about containers: one dealing with loading solid objects into boxes, the other dealing with carrying liquids in open containers and pouring from one container into another. He has written a textbook on representations of commonsense knowledge (Davis, 1990).

Gary Marcus is a cognitive scientist in the psychology department at NYU. He has published experimental and theoretical work in *Science, Nature, Cognition, Cognitive Psychology*, and numerous other leading journals. He is also the editor of *The Norton Psychology Reader*, and author of four books about cognitive science, including the *New York Times* bestseller *Guitar Zero: The New Musician and The Science of Becoming Musical* and *Kluge: The Haphazard Construction of the Human Mind*, which was a *New York Times Book Review* editor's choice.

Since September 2012 Davis and Marcus have been collaborating on a number of projects. They have written a critique of Bayesian methods of high-level cognition), and are currently writing a paper on the limits of simulation in cognitive models and models of automated reasoning.

8.4 Institutional support

Davis' research is supported by the NYU Computer Science department and the Courant Institute of Mathematical Sciences. Marcus' research is supported by the NYU Psychology department and the

School of Arts and Science. All aspects of the administrative, scientific, and computational infrastructure required by this project receive full support from these parts of the university.

9. Broader Impact

1. **Training of graduate students**. The proposed budget contains 1 semester of academic year support and two summer months of support per year for one student in computer science and one student in psychology. Over Davis' career he has supervised nine doctoral students, who are now working at such places as IBM Watson Labs, Microsoft Research, the Hebrew University, and ISI; Marcus's two recent PhD students are currently post-docs at Harvard and Toronto.

2. **Public understanding of science**.

a. Writing: The Co-PI (Marcus) frequently writes about complex, technical subjects for the general-public, mostly recently in an ongoing series of widely-read essays on artificial intelligence and cognitive science at the website of *The New Yorker*. He has also written for *The New York Times* (both for the *Science Times* and *Sunday Magazine*), *Discover*, *Wired*, and the *Wall St Journal*, and envisions one or more lay-audience articles in prominent places on the topic of common-sense reasoning, informed by the present work.

b. Video: Additionally, in conjunction with a new collaborator, the filmmaker Jason Silva (whose work recently opened TEDGlobal, Marcus aims to develop a video (presumably for free distribution on YouTube) comparing human and machine common-sense reasoning.

c. Public Appearances: Marcus will describe our work and broader issues that arises in common-sense reasoning, in public lectures and radio appearances (he has spoken at venues such as TED NYC and appeared numerous times on NPR and other radio shows).

d. Song: The well-known rapper and playwright Dirk Murray "Baba" Brinkman is interested in the possibility of writing a rap song on the subject of AI and commonsense reasoning in AI. Brinkman's 2011 show, "A Rap Guide to Evolution" was extremely successful and received a glowing review in the New York TImes. (For this subproject, we would seek outside funding, rather than NSF fnds.)

e. Museum exhibit: We hope to help develop a museum exhibit on AI and Common Sense Reasoning, and are in discussions with Paul Hoffman, President and CEO of Liberty Science Museum, who has expressed interest. (Marcus has also close ties to the American Museum of Natural History; he has consulted on the development of two exhibits, and appears in a video that is shown regularly there in The Hall of Human Origins).

References

W. Benger, "Colliding galaxies, rotating neutron stars and merging black holes—visualizing high dimensional datasets on arbitrary meshes." *New Journal of Physics,* **10,** 2008. http://iopscience.iop.org/1367-2630/10/12/125004

D.G. Bobrow (ed.), *Qualitative Reasoning about Physical Systems*, MIT Press, 1985.

G.Brewka, I. Nielmelä, and M. Truszczński, "Nonmonotonic Reasoning" in F. van Harmelen, V. Lifschitz, and B. Porter (eds.) *Handbook of Knowledge Representation*, Elsevier, 2008, 239-284. .

M. Christani, “The complexity of reasoning about spatial congruence,” *Journal of Artificial Intelligence Research,* **11**, 1999, 361-390. http://www.jair.org/media/641/live-641-1837-jair.pdf

A.G. Cohn and J. Renz, “Qualitative Spatial Representation and Reasoning.” In F. van Harmelen, V. Lifschitz, and B. Porter (eds.) *Handbook of Knowledge Representation*, Elsevier, 2008, 551-596.

B. Cook, D. Kroening, and N. Sharygina, “Accurate Theorem Proving for Program Verification,” in T. Magaria and T. Steffen (eds.) *ISoLA 2004*, LNCS 4313, Springer, 2006, 96-114. http://www.inf.usi.ch/faculty/sharygina/cogent.pdf

M. Cox and A. Raja, *Metareasoning: Thinking about Thinking*, MIT Press, 2011.

E. Davis, *Representations of Commonsense Knowledge*, Morgan Kaufmann, 1990.

E. Davis, “Pouring Liquids: A Study in Commonsense Physical Reasoning,” *Artificial Intelligence*, **172**, 2008, 1540-1578. http://www.cs.nyu.edu/faculty/davise/papers/liquids.pdf

E. Davis, “How Does a Box Work? A Study in the Qualitative Dynamics of Solid Objects”, *Artificial Intelligence*, **175**, 2011, 299-345. http://www.sciencedirect.com/science/article/pii/S0004370210000421

E. Davis, “Qualitative Spatial Reasoning in Interpreting Text and Narrative,” *Spatial Cognition and Computation,* 2013. http://www.cs.nyu.edu/faculty/davise/papers/cosit.pdf

E. Davis and G. Marcus, “The Limits of Simulation as an AI and a Cognitive Theory of High-Level Physical Reasoning,” journal paper in prep.

J. de Kleer and J.S. Brown, "A Qualitative Physics based on Confluences," in D. Bobrow (ed.) *Qualitative Reasoning about Physical Systems*, MIT Press, 1985.

M.Egenhofer and R. Franzosa, “Point-set topological spatial relations*,” International Journal of Geographical Information Systems*, **5**(2), 1991, 161-174. ftp://ftp.geoinfo.tuwien.ac.at/wilke/BUP_Skriptsammlungen/GeoInfo/[Pub]/Egenhofer/[Egenhofer+Franzosa]_Pointset-Topological_Spatial_Relations.pdf

K. Forbus, "Qualitative Process Theory," in D. Bobrow (ed.) *Qualitative Reasoning about Physical Systems*, MIT Press, 1985.

K. Forbus, "Qualitative Modeling", *Wiley Interdisciplinary Reviews: Cognitive Science*, **2**(4), 2011, 374-391.

K.Fo rbus and J. de Kleer, *Building Problem Solvers*, MIT Press, 1993.

B. Funt, Problem-solving with Diagrammatic Representations, *Artificial Intelligence*, **13**(3), 1980, 201-230,

F. Gardin and B. Meltzer, “Analogical Representations of Naïve Physics,” *Artificial Intelligence*, **38**, 1989, 139-159. http://www.sciencedirect.com/science/article/pii/0004370289900556

M. Gelfond, “Answer Sets”. In F. van Harmelen, V. Lifschitz, and B. Porter (eds.) *Handbook of Knowledge Representation*, Elsevier, 2008, 285-316.

C.P. Gomes, “Satisfiability Solvers”. ”. In F. van Harmelen, V. Lifschitz, and B. Porter (eds.) *Handbook of Knowledge Representation*, Elsevier, 2008 89-134.

J. Hamrick, P. Battaglia, and J. Tenenbaum, "Internal Physics Models Guide Probabilistic Judgement about Object Dynamics", *Proc. 33rd Annual Conf. of the Cognitive Science Society*, 2011.

H. Hecht and D.R. Proffitt. "The price of expertise: Effects of experience on the water-level task." *Psychological Science*, **6**(2), 1995, 90-95.

B.Y. Inhelder, and J. Piaget, *The growth of logical thinking from childhood to adolescence.* New York: Basic Books, 1958.

A. Joh, V.K. Jaswal and R. Keen, "Imagining a Way Out of the Gravity Bias: Preschoolers Can Visualize the Solution to a Spatial Problem," *Child Development*, **82**(3), 2011, 744-750.

B. Johnston, *Practical Artificial Commonsense*, Ph.D. thesis, University of Technology, Sydney, 2010.

H. Kaufmann and B. Meyer, "Simulating educational physical experiments in augmented reality", SIGGRAPH Asia, 2008.

B. Kuipers, "Qualitative Simulation", in D. Bobrow (ed.) *Qualitative Reasoning about Physical Systems*, MIT Press, 1985.

V. Lifschitz, "Cracking an Egg: An Exercise in Commonsense Reasoning." *Logical Formalizations of Commonsense Reasoning*, 1998,

G. Marcus and E. Davis, "Probabilistic models of higher-level cognition and the dangers of confirmationism," submitted to *Psychological Science*.

K.D. Markman, W.M. Klein, and J. Suhr, *Handbook of Imagination and Mental Simulation*, Psychology Books, 2009.

M. McCloskey, "Intuitive physics." *Scientific American,* **248**(4), 1983, 114-122.

W. McCune, "Prover9 and Mace4", http://www.cs.unm.edu/~mccune/mace4/

S. Murman et al., "An Interface for Specifying Rigid-Body Motions for CFD Applications," *AIAA*, 2003. http://people.nas.nasa.gov/~smurman/AIAA2003-1237.pdf

L. Morgenstern, "Mid-Sized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking," *Studia Logica*, 67(3):333-384, 2001. http://www.jstor.org/stable/20016288

A. Rahimian et al. "Petascale Direct Numerical Simulation of Blood Flow on 200K Cores and Heterogeneous Architectures." *Supercomputing* 2010, 1-11. http://dl.acm.org/citation.cfm?id=1884648

D.A. Randell, Z. Cui, and A.G. Cohn. A spatial logic based on regions and connection. *Third International Conference on Principles of Knowledge Representation and Reasoning.* 1992. 165-176.

M. Shanahan, "An Attempt to Formalise a Non-Trivial Benchmark Problem in Common Sense Reasoning." *Artificial Intelligence*, 153(1-2): 2001, 141-165, http://www-formal.stanford.edu/leora/shanahan-aij-egg.ps

M. Siegal, *Knowing children: Experiments in conversation and cognition.* Mahwah NJ: Elrbaum, 1991.

A. Tabiei and G. Nilakantan. "Reduction of Acceleration Induced Injuries from Mine Blasts under Infantry Vehicles." Dept. of Aerospace Engineering and Engineering Mechanics, undated http://www.ase.uc.edu/~atabiei/pdfs/J.pdf

E. Téglás et al. "Intuitions of probabilities shape expectations about the future at 12 months and beyond," *Proc. Nat. Academy of Sciences*, **104**(48) 2007, 19156-19159. http://www.ncbi.nlm.nih.gov/pubmed/18025482

E. Téglás et al., "Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference," *Science*, **332**, 2011, 1054-1059. http://www.sciencemag.org/content/332/6033/1054.short

K. Velten, *Mathematical Modeling and Simulation*, Wiley, 2009.

C. Weidenbach et al. "SPASS Version 2.0", *CADE-18*, *Lecture Notes in Computer Science*, *2392,* Springer, 2002, 275-279. http://dl.acm.org/citation.cfm?id=757359

S. Wölfl, T. Mossakowski, and L. Schröder, "Qualitative Constraint Calculi: Heterogeneous Verification of Composition Tables", *AAAI,* 2007. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.8824

S. Zickler and M. Veloso, "Efficient Physics-Based Planning: Sampling Search via Non-Deterministic Tactics and Skills", *8th Int. Conf. on Autonomous Agents and Multiagent Systems*, 2009, 27-34. http://dl.acm.org/citation.cfm?id=1558015