Constructing and Refining Causal Explanations from an Inconsistent Domain Theory [1]

Richard J. Doyle

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Recent work in the field of machine learning has demonstrated the utility of explanation formation as a guide to generalization. Most of these investigations have concentrated on the formation of explanations from consistent domain theories. I present an approach to forming explanations from domain theories which are inconsistent due to the presence of abstractions which suppress potentially relevant detail. In this approach, explanations are constructed to support reasoning tasks and are refined in a failure-driven manner. The elaboration of explanations is guided by the structuring of domain theories into layers of abstractions.

This work is part of a larger effort to develop a causal modelling system which forms explanations of the underlying causal relations in physical systems. This system utilizes an inconsistent, common-sense theory of the mechanisms which operate in physical systems.

1 The Problem

The field of machine learning has shown a recent shift towards knowledge intensive methods which utilize the construction of explanations as an important step in the generalization process.

In these *explanation-based learning* methods [DeJong & Mooney 86, Mahadevan 85, Mitchell et al 86, Winston et al 83], an explanation derived from a domain theory shows why a particular example is an instance of some concept. After the critical constraints in the explanation are determined, its components are generalized while maintaining these constraints; the result is a generalized recognition rule for examples of the given concept.

This approach is now well understood for domain theories which are *consistent*, or are at least assumed to be consistent. Explanations derived and generalized from consistent domain theories constitute *proofs* which can be taken to be correct in the context of all reasoning tasks they may subsequently support.

However, most domain theories are not consistent – they incorporate defaults, they omit details, or they otherwise abstract away from a complete account of the constraints which may be relevant to the reasoning tasks to which they are applied. Explanations derived and generalized from inconsistent domain theories cannot be assumed to be always correct; their inherent abstractions may manifest when inferences derived from them are not corroborated.

The problem addressed in this paper is how to construct justified, plausible explanations despite inconsistent domain theories, and how to refine those explanations or their generalizations when they fail to support reasoning tasks to which they are applied.

1.1 An Example

Consider a domain theory which describes at a common-sense level the kinds of causal mechanisms that operate in physical systems: flows, mechanical couplings, etc. My system derives from this domain theory a simple causal model of a bathtub which describes two flow mechanism instances: water flows in at the tap and flows out at the drain. This simple model proves inadequate for the planning problem of how to fill the bathtub with water. This reasoning task becomes solvable after my system elaborates the model to describe a mechanical coupling between the plunger and the plug and how the plug blocks the flow of water at the drain.

This elaborated causal explanation includes an interesting intersection between a flow mechanism and a mechanical coupling mechanism. A single physical object – the plug – plays dual roles: it serves both as one half of a mechanical coupling and as barrier to a flow. My system extracts this composed causal mechanism – which might be called "valve" – out of the causal model of the bathtub and generalizes it in the explanation-based learning manner, maintaining the constraint that one physical object play these two roles.

My system next uses the valve mechanism to explain the causal relations in another physical system – a camera. In a camera, there is a mechanical coupling between the shutter release and the shutter; furthermore the shutter plays the additional role of barrier to the light flow between the photographed subject and the film. This causal model generates an incorrect prediction when a lens cap is inadvertently left on the lens. My system refines the model by instantiating the lens cap as another barrier to light flow. The model also cannot be used to explain why the shutter does not move when the safety latch on the shutter release is engaged. My system handles this situation by instantiating a latch as a type of barrier to a mechanical coupling

[1] This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N000414-80-C-0505.

– a detail which never appeared in the original construction of the valve explanation in the context of the bathtub.

1.2 The Proposed Solution

I take the following view of explanation formation from inconsistent domain theories, as a tool for learning or otherwise: Explanations are constructed in the context of a reasoning task; they are refined, as needed, in an incremental failure-driven manner. The usefulness of an explanation is relative to the goal of its motivating reasoning task. Similarly, the consistency of an explanation is relative to the set of inferences it supports.

In the example above, a planning problem motivates the elaboration of the bathtub model and prediction failures motivate the refinement of the camera model.

In this paper, I focus on domain theories which are inconsistent because they incorporate a particular kind of abstraction – the suppression of possibly relevant detail through *approximation*. Approximations may be layered into several levels. Explanations derived from less approximate levels are less likely to support incorrect inferences.

I argue that the minimum level to which an explanation must be instantiated depends on the goal of the motivating reasoning task. I present two means of refining failed explanations: reinstantiation of the explanation into a situation which has changed, and elaboration of the explanation to a less approximate level in the domain theory with more explanatory power. This approach to refinement uses the layered structure of a domain theory to guide the familiar processes of dependency-directed backtracking and truth maintenance [Doyle 79].

2 A Context for the Problem – Causal Modelling

The issue of how to construct and refine explanations from an inconsistent domain theory comes up in my work on *causal modelling* [Doyle 86]. My causal modelling system learns how physical systems work in the context of reasoning tasks such as planning or prediction. Given a description of how quantity values, structural relations, and geometrical relations in a physical system change over time, my system utilizes a common-sense theory of causal *mechanisms* to hypothesize underlying causal relations which can explain the observed behavior of the physical system.

I have developed a representation for causality in physical systems which supports the description of these mechanisms or processes by which effects emerge from causes in this domain. This aspect of my work addresses issues first considered in [Rieger & Grinberg 77].

In my representation, mechanisms require the presence of some kind of *medium*, or structural link, between the site of the cause and the site of the effect. For example, flows require a channel through which to transfer material and mechanical couplings require a physical connection through which to propagate momentum. Causal mechanisms can be disrupted by *barriers* which decouple cause from effect. For example, flows can be inhibited by a blocked channel and mechanical couplings can be disabled by a broken physical connection.

This representation for causality and a vocabulary of causal mechanisms describable within it are currently under development and are being tested in the modelling of a number of physical systems. A generalization hierarchy for these mechanisms is shown in Figure 1.

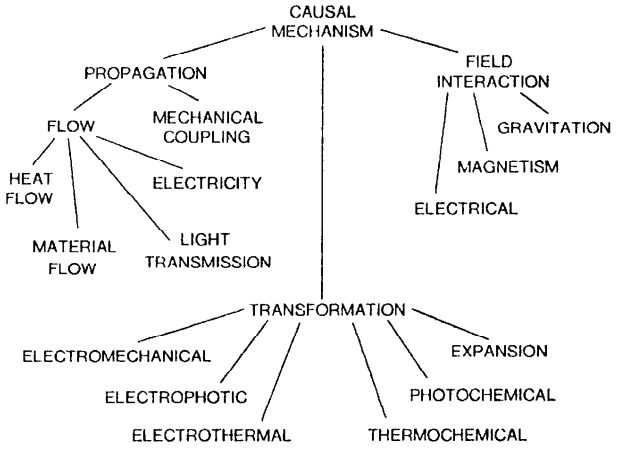

Figure 1: Generalization Hierarchy for Causal Mechanisms

The relevant aspect of this domain theory of causal mechanisms for the purposes of this paper is its inconsistency. The theory does not describe all the relevant aspects of the various mechanisms which operate in physical systems. Furthermore, the representation of causality underlying this domain theory suggests a decomposition of the mechanism descriptions into several layers of approximation. I describe these levels of explanation in the next section.

2.1 Layers of Explanation in a Domain Theory

There are several levels of causal explanation available in the representation for causality described above, each drawing on the notion of mechanism to a different degree. Each more detailed level introduces additional constraints which are meaningful only in the context provided by the more abstract levels. The higher levels of explanation do not employ a coarser grain size, rather they ignore certain potentially relevant conditions.

The most abstract level of explanation in the representation does not incorporate the notion of causal mechanism at all. This explanation merely notes the co-occurrence of two events and verifies that the effect does not precede the cause.

CO-OCCURRENCE EXPLANATION

$$\exists(iq)\exists(dq)$$
$$(Changes(iq, t1) \wedge Changes(dq, t2) \quad \wedge \quad \geq (t1, t2))$$
$$\Longrightarrow$$
$$FunctionalDependence(iq, dq)$$

The next level of explanation verifies that the quantities whose values are correlated are of the appropriate type for the mechanism. For example, flows are causal links between *amount* quantities and mechanical couplings are causal links between *position* quantities. [2]

QUANTITY TYPES EXPLANATION

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IndependentQuantityType(iq) \wedge DependentQuantityType(dq))$
$\Longrightarrow$
$FunctionalDependence(iq,dq)$

This explanation is an approximation of one which identifies the enabling medium between the physical objects of the quantities whose values are correlated.

MEDIUM EXPLANATION

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IndependentQuantityType(iq) \wedge$
$DependentQuantityType(dq) \wedge$
$\exists(m)$
$(Between(m, PhysicalObjectOf(iq), PhysicalObjectOf(dq),$
$t1:t2) \wedge MediumType(m)))$
$\Longrightarrow$
$FunctionalDependence(iq,dq)$
$Enables(m, FunctionalDependence(iq,dq))$

Note that the medium must be maintained throughout the causal interaction.

This explanation in turn approximates one which states that there must be no barriers which disrupt the structural link and disable the causal mechanism.

BARRIER EXPLANATION

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge \neg Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IndependentQuantityType(iq) \wedge$
$DependentQuantityType(dq) \wedge$
$\exists(m)$
$(Between(m, PhysicalObjectOf(iq), PhysicalObjectOf(dq),$
$t1:t2) \wedge MediumType(m) \wedge$
$\exists(b)$
$(Along(b,m,t1:t2) \wedge BarrierType(b))))$
$\Longrightarrow$
$FunctionalDependence(iq,dq)$
$Enables(m, FunctionalDependence(iq,dq))$
$Disables(b, FunctionalDependence(iq,dq))$

Finally, this description of barriers can be elaborated to one which states that in general the effectiveness of a barrier depends on how much of the medium it blocks.

VARIABLE BARRIER EXPLANATION [3]

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IndependentQuantityType(iq) \wedge$
$DependentQuantityType(dq) \wedge$
$\exists(m)$
$(Between(m, PhysicalObjectOf(iq), PhysicalObjectOf(dq),$
$t1:t2) \wedge MediumType(m) \wedge$
$\exists(b)$
$(Along(b,m,t1:t2) \wedge BarrierType(b) \wedge$
$\exists(bq)$
$(QuantityOf(bq,b) \wedge IsA(bq, Position)))))$
$\Longrightarrow$
$FunctionalDependence(iq,dq)$
$Enables(m, FunctionalDependence(iq,dq))$
$FunctionalDependence(bq,dq)$
$Enables(b, FunctionalDependence(bq,dq))$

Note that this level of explanation describes a dependence (between a quantity associated with a barrier and the quantity associated with the effect) which does not appear at any of the other levels.

These levels of explanation are defined for causal mechanisms in general; the particular most detailed levels of explanation of flows and mechanical couplings needed for the bathtub and camera examples are shown below.

MATERIAL FLOW (VARIABLE BARRIER)

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IsA(iq, Amount) \wedge IsA(dq, Amount) \wedge$
$\exists(m)$
$(Touches(PhysicalObjectOf(iq), PhysicalObjectOf(dq),$
$t1:t2) \wedge$
$\exists(b)$
$(Along(b,m,t1:t2) \wedge Blocks(b, PhysicalObjectOf(iq)) \wedge$
$\exists(bq)$
$(QuantityOf(bq,b) \wedge IsA(bq, Position)))))$
$\Longrightarrow$
$FunctionalDependence(iq,dq)$
$Enables(m, FunctionalDependence(iq,dq))$
$FunctionalDependence(bq,dq)$
$Enables(b, FunctionalDependence(bq,dq))$

LIGHT TRANSMISSION (VARIABLE BARRIER)

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IsA(iq, Amount) \wedge IsA(dq, Amount) \wedge$
$\exists(m)$
$(StraightLinePath(PhysicalObjectOf(iq),$
$PhysicalObjectOf(dq), t1:t2) \wedge$
$\exists(b)$
$(Along(b,m,t1:t2) \wedge Opaque(b) \wedge$
$\exists(bq)$
$(QuantityOf(bq,b) \wedge IsA(bq, Position)))))$
$\Longrightarrow$
$FunctionalDependence(iq,dq)$
$Enables(m, FunctionalDependence(iq,dq))$
$FunctionalDependence(bq,dq)$
$Enables(b, FunctionalDependence(bq,dq))$

MECHANICAL COUPLING (BARRIER)

$\exists(iq)\exists(dq)$
$(Changes(iq,t1) \wedge \neg Changes(dq,t2) \wedge \geq(t1,t2) \wedge$
$IsA(iq, Position) \wedge IsA(dq, Position) \wedge$
$\exists(m)$
$(AttachedTo(PhysicalObjectOf(iq), PhysicalObjectOf(dq),$
$t1:t2) \wedge$

[2] Flows and mechanical couplings are instances of a class of causal mechanisms I call *propagations*; they involve similar co-occurring events at different sites. There are also *transformations* (e.g. photochemical on film, electrophotic in a light bulb, electrothermal in a toaster) involving different co-occurring events at a single site.

[3] This level of explanation removes a different type of abstraction than the other levels. This difference is discussed in the section on types of abstraction.

$\exists(b)$
$(AttachedTo(b, m) \wedge Anchored(b))))$
$\Longrightarrow$
$FunctionalDependence(iq, dq)$
$Enables(m, FunctionalDependence(iq, dq))$
$Disables(b, FunctionalDependence(iq, dq))$

This last explanation describes a latch barrier to a mechanical coupling.

Although this presentation of levels in the causal mechanism descriptions suggests fixed approximation hierarchies, in general there may be several ways to elaborate any level of explanation. For example, a non-rigid physical connection as well as a latch may disable a mechanical coupling.

3 Constructing and Refining Explanations

In this section, I show how the causal explanations of the bathtub and camera are constructed and incrementally refined from the layered, approximate, inconsistent domain theory described in the previous section.

3.1 Construction of an Explanation

Consider the problem of constructing a causal model of a bathtub in the context of a planning problem to fill a tub with water. A first attempt instantiates a model at the medium level of explanation of the material flow mechanism. This causal explanation is shown in Figure 2. [4]

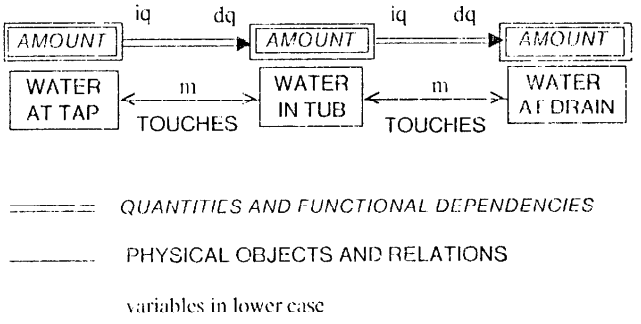

Figure 2: Medium Level Explanation of Material Flows in a Bathtub

Water flows from the tap through the tub and out of the drain.

It is not yet possible to generate a plan for filling the tub with water. Starting a flow at the tap *into* the tub does not work because of the presence of the drain, an additional medium for flow *from* the tub. This unsolved planning problem motivates the further expansion of the bathtub model to a more detailed level of explanation, as shown in Figure 3.

The model now reveals how the plug can block flow at the drain and how the position of the plug is affected by the plunger.

[4] Temporal, enabling, and disabling relations are mostly suppressed in the figures to avoid clutter; such information is used as indicated in the explanation descriptions.

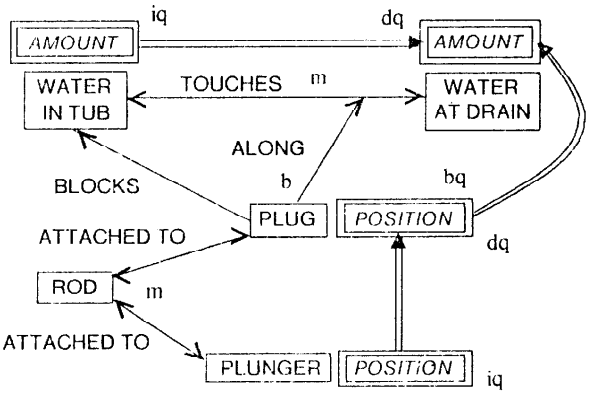

Figure 3: Variable Barrier Level Explanation of a Material Flow and Medium Level Explanation of a Mechanical Coupling in a Bathtub

A plan can now be generated for filling the bathtub with water: Start a flow of water at the tap and move the plunger to place the plug in the drain.

This example of explanation construction illustrates how the needed level of explanation depends on the motivating reasoning task. In this case, a barrier was needed to solve the planning problem; hence the causal explanation of the bathtub had to go to the barrier level.

3.2 Generalization of an Explanation

A material flow mechanism and a mechanical coupling mechanism intersect in the expanded bathtub model at the plug. My causal modelling system notes such intersections because they may provide opportunities for extracting and generalizing useful compositions of causal mechanisms. This particular complex mechanism might be called "valve".

Using the hierarchy in Figure 1, my system generalizes the valve concept to other kinds of flows. The definitions of media, barriers, etc. for other types of flow are substituted while maintaining the constraint that a single physical object must serve as both one half of the mechanical coupling and as barrier to the flow. The generalized valve mechanism for light transmission is shown in Figure 4.

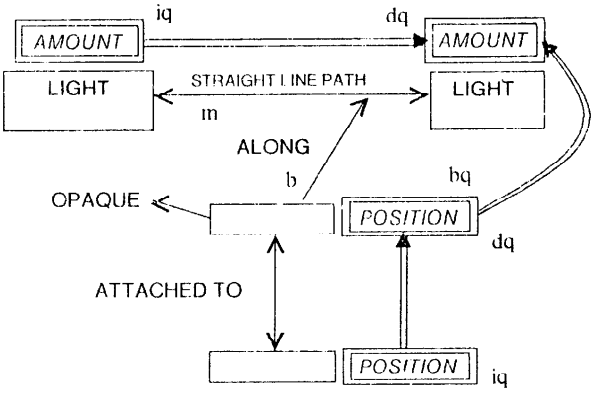

Figure 4: The Learned Valve Mechanism for Light Transmission

This learned complex mechanism is used in the construction of a causal model for another physical system - a camera. This causal model is shown in Figure 5.

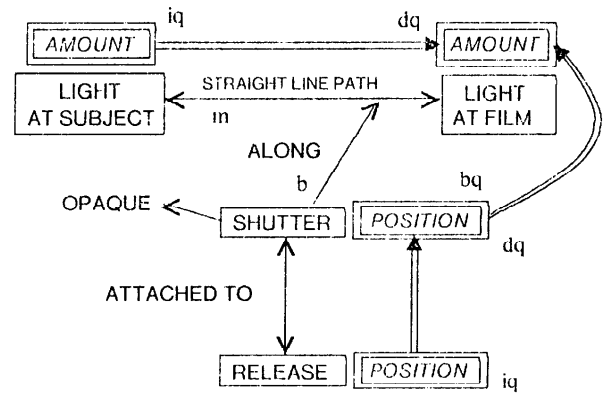

Figure 5: Valve Explanation of a Camera

All of the valve explanations combine the variable barrier explanation level of a flow mechanism and the medium explanation level of the mechanical coupling mechanism. The origins of composed mechanism explanations are recorded, as in Figure 6, so that more detailed levels of explanation in the constituent mechanisms can be accessed if needed.

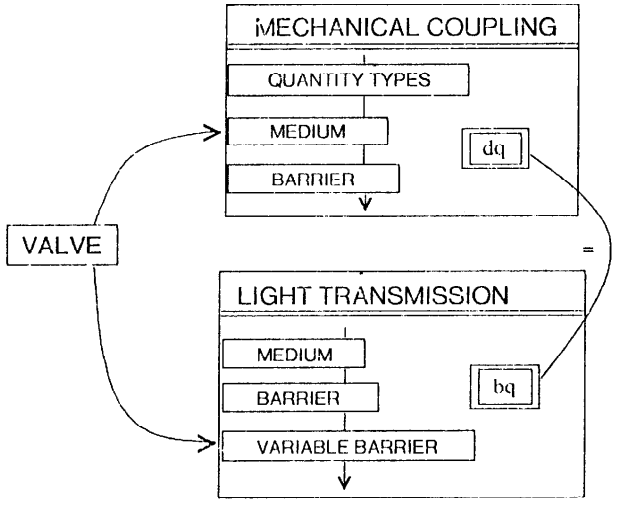

Figure 6: Origins of the Valve Explanation for Light Transmission

3.3 Refinement of an Explanation through Reinstantiation

When a lens cap is placed on a camera this model supports an incorrect prediction - that light will continue to reach the film. In this case, the level of explanation needed to handle the new situation already appears in the model; the lens cap, like the shutter, is a barrier to light flow. My system instantiates this additional barrier, as in Figure 7. The refined explanation now supports the correct prediction that light does not reach the film in the altered camera.

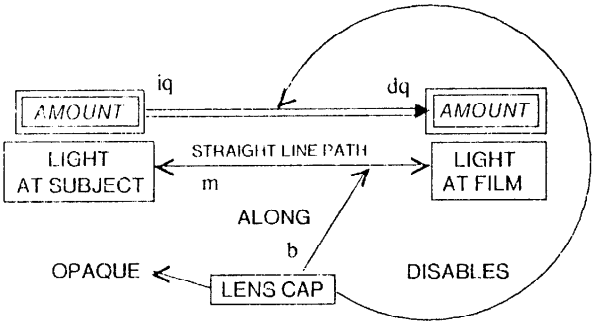

Figure 7: Reinstantiated Explanation of a Camera

3.4 Refinement of an Explanation through Elaboration

In some cases, refinement of a failed explanation requires elaborating to a level of explanation which calls on details not yet considered. This kind of refinement is needed in the camera model to handle the situation where a safety latch on the shutter release is engaged. As is, the model supports the incorrect inference that the shutter moves whenever the release moves.

The model is repaired when my system recognizes the latch barrier to the mechanical coupling between the release and the shutter, as in Figure 8. The shutter does not move when the anchored release latch is attached. My system formed this explanation by elaborating to the barrier explanation level of the mechanical coupling constituent of the valve mechanism for light transmission (see Figure 6). Although this level of explanation was never reached in the bathtub model, it is accessible in the learned valve mechanism for light transmission used in the camera model.

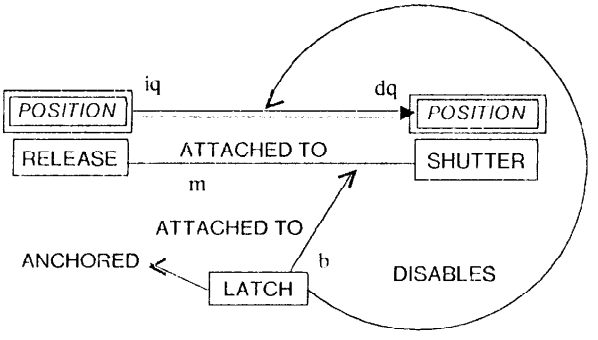

Figure 8: Elaborated Explanation of a Camera

4 Issues

In this section, I discuss a set of issues relevant to the problem of constructing and refining explanations from a domain theory which is inconsistent.

4.1 Limits on Perception and Use of Empirical Evidence

The *justification* for employing an explanation to support reasoning in a given situation comes partially from the explanation schema used, and partially from the *perceptions* which instantiate the existentially quantified terms in that explanation schema. The justification due to the explanation schema may be compromised by approximations. The justification due to perception may be compromised when the instantiations of terms in an explanation are *unobservable* due to limits in the available perception equipment.

For example, air, which may be unobservable, serves as a thermal conducting medium for heat flow in a toaster. A causal explanation for a toaster based on heat flow might be only partially instantiated.

The loss of justification due to an uninstantiable term can be countered by gathering *empirical* evidence that an explanation is consistent, e.g., confirming that bread placed in a toaster does indeed become hotter. This is one way in which analytical, i.e., explanation-based, methods can be combined with empirical methods.

Uninstantiable terms in an explanation also can be countered by elaborating an explanation. More detailed levels of explanation can suggest how to obtain indirect empirical evidence for the uninstantiable term. For example, the barrier level of explanation in the heat flow mechanism indicates that heat flow in a toaster should be disabled when a thermal insulator exists between the coils and the bread. Confirming observations at this level can strongly suggest the presence of the unobservable thermal conducting medium.

4.2 Types of Abstraction

Approximation is the most prevalent type of abstraction appearing between levels of explanation in the causal mechanism domain theory. Approximations are assumptions that some condition holds or that some constraint is satisfied. For example, the approximation between the quantity types and medium levels of explanation is that an appropriate medium to support a causal mechanism is in place. A more detailed explanation may be correct in situations where an approximate explanation is not.

A different kind of abstraction appears between the barrier and variable barrier levels. Here a continuous description is collapsed into a discrete one. At the barrier level, a barrier either completely disables a mechanism or has no effect at all. At the variable barrier level a barrier may also partially affect a mechanism. This kind of abstraction might be called *qualitization*. Some situations may not even be describable, much less correctly described, by explanations which incorporate qualitizations. For example, the variable flow out of a bathtub drain or the way the aperture in a camera lens affects light flow cannot be described by the on/off barrier explanation.

Under *aggregation*, complex structures at one level of explanation are subsumed under simpler structures, perhaps even single terms, at higher levels. Aggregations involve changes in grain size. The oft-used example is the alternate explanations of gas behavior in terms of the motions of molecules and in terms of the macroscopic properties of volume, temperature, and pressure. Aggregations currently do not appear in the causal mechanism domain theory; the theory stops short of a full physical accounting of the laws which govern the behavior of physical systems.

This enumeration of abstraction types is admittedly preliminary. A recent investigation [Smith et al 85] also has described different abstraction types, and has investigated how explanations fail because of them. I have described in this paper an approach to explanation construction and refinement from domain theories which incorporate approximations and qualitizations.

4.3 Incomplete and Intractable Domain Theories

Even the lowest level of explanation in a domain theory may incorporate abstractions. This is true of the causal mechanism theory. For example, a barrier may be selective, e.g. a UV filter on a camera. Abstractions at the lowest level of a domain theory imply missing knowledge.

The method of explanation refinement described in this paper has no recourse when an incomplete domain theory "bottoms out". A possible course of action in this circumstance is to resort to an inductive method. Another is to invoke some other means of accessing applicable knowledge, perhaps analogy. Simply giving up may also be arguably appropriate.

Even complete domain theories might make use of layered approximations. A complete domain theory may involve so much detail as to be intractable. A structuring of such a theory into several approximating levels of explanation allows plausible explanations to be constructed, and maintains a capability for refining those explanations [Tadepalli 85].

4.4 Learning from Experiments

Given an inconsistent domain theory, it is possible to derive more than one plausible, partially justified explanation in many situations. For example, a glowing taillight on an automobile might be explained either by the electrical system of the car or by reflected light from the sun.

I am developing an experiment design capability for distinguishing multiple explanations. This capability utilizes the explanation refinement method described in this paper. It appears similar in spirit to that proposed in [Rajamoney et al 85]. In my method, refinements are proposed to one or more of a set of competing explanations until the explanations support divergent predictions. The refinements specify further instantiations at the same or at an elaborated level. For example, an experiment to distinguish the glowing taillight explanations might elaborate the light flow explanation from the medium level and specify the instantiation of an opaque barrier to disable the hypothesized light transmission. This barrier, importantly, would have no predicted effect on the electrical system of the car.

This approach to experiment design applies equally well to single explanations. Even an explanation with no rivals may be only partially justified because of perception limits. Empirical evidence for the correctness of such an explanation may be gathered via experiments which specify refinements involving addi-

tional *observable* instantiations of terms in the explanation. Such an experiment, involving a toaster, is described in the section on perception above.

Experimenting can be viewed as the active gathering of greater justification for fewer and fewer plausible explanations.

5 Relation to Other Work

Patil has investigated multi-level causal explanation in a medical domain [Patil et al 81]. He identifies five levels of explanation and describes methods for moving between levels in both directions. The kind of abstraction employed by Patil's system ABEL is aggregation; nodes and/or causal links at one level are condensed into fewer nodes and links at the next higher level. Elaboration in ABEL supports confirmation of diagnoses to greater resolution and allows the reasoning of the system to be revealed in greater detail to a user. Elaboration in ABEL is not intended to support failure-driven refinement of explanations through the removal of approximations, as described in this paper.

Davis' hardware troubleshooting system expands both aggregations and approximations [Davis 84]. The structure and behavior of digital circuits are described at several levels of aggregation; this provides the troubleshooting system with different grain sizes at which to examine a circuit. *Fault models* indicate how to lift approximations concerning the possible "paths of interaction" in circuits. Davis' fault models appear to be well-described in my representation for causality. His notions of spurious and inhibited causal pathways correspond to my concepts of medium and barrier.

In general, there may be many ways to repair failed approximate explanations. Smith et al [Smith et al 85] have explored how the task of isolating the source of an explanation failure can be constrained. They show how different types of abstraction in an explanation schema propagate along dependency links to instantiated explanations and lead to different types of failure.

6 Conclusions

I have presented an approach to explanation construction and refinement from inconsistent domain theories which incorporate two types of abstraction – the suppression of potentially relevant constraints and the discretization of continuous representations. In this approach, explanations are elaborated to support new reasoning tasks and to recover from failures. The elaboration process is guided by the structuring of domain theories into layers of abstractions.

This work is taking place in the context of an investigation into the formation of causal models of physical systems. Causal modelling involves the construction and refinement of causal explanations of the behavior of physical systems from a domain theory describing the mechanisms which operate in such systems. The levels of explanation in this domain theory are derived from a representation for causality in physical systems.

Some of the issues related to explanation formation from inconsistent domain theories include: using empirical evidence to complement explanation, understanding the types of abstraction which render a domain theory inconsistent, dealing with the incompleteness of a domain theory, and designing experiments to distinguish and gather justification for plausible explanations. In addition, a better understanding is needed of the kinds of domain theories which admit to decomposition via layered abstractions, and of the principles which govern the placement of orderings on abstractions.

7 Acknowledgements

Patrick Winston encouraged me to pursue this line of investigation. Randall Davis, Bob Hall, David Kirsh, Rick Lathrop, Jintae Lee and Tomas Lozano-Perez have all engaged in discussions of the ideas in this paper.

8 References

[Davis 84] Davis, Randall, "Diagnostic Reasoning Based on Structure and Behavior," *Artificial Intelligence* **24**, 1984.

[DeJong & Mooney 86] DeJong, Gerald and Raymond Mooney, "Explanation Based Learning: A Differentiating View," *Machine Learning* **1**, no. 2, 1986.

[Doyle 79] Doyle, Jon, "A Truth Maintenance System," *Artificial Intelligence* **12**, 1979.

[Doyle 86] Doyle, Richard J., "Construction and Refinement of Justified Causal Models Through Multiple Levels of Explanation and Experimenting," Ph.D., Massachusetts Institute of Technology, *forthcoming.*

[Mahadevan 85] Mahadevan, Sridhar, "Verification-Based Learning: A Generalization Strategy for Problem-Reduction Methods," *9th IJCAI*, 616-623, 1985.

[Mitchell et al 86] Mitchell, Tom M., Richard M. Keller and Smadar T. Kedar-Cabelli, "Explanation-Based Generalization: A Unifying View," *Machine Learning* **1**, no. 1, 1986.

[Patil et al 81] Patil, Ramesh S., Peter Szolovits and William B. Schwartz, "Causal Understanding of Patient Illness in Medical Diagnosis," *7th IJCAI*, 1981.

[Rajamoney et al 85] Rajamoney, Shankar, Gerald DeJong and Boi Faltings, "Towards a Model of Conceptual Knowledge Acquisition Through Directed Experimentation," *9th IJCAI*, 688-690, 1985.

[Rieger & Grinberg 77] Rieger, Chuck and Milt Grinberg, "The Declarative Representation and Procedural Simulation of Causality in Physical Mechanisms," *5th IJCAI*, 1977.

[Smith et al 85] Smith, Reid G., Howard A. Winston, Tom M. Mitchell and Bruce G. Buchanan, "Representation and Use of Explicit Justifications for Knowledge Base Refinement." *9th IJCAI*, 673-680, 1985.

[Tadepalli 85] Tadepalli, Prasad V., "Learning in Intractable Domains," *3rd International Machine Learning Workshop*, 202-205, 1985.

[Winston et al 83] Winston, Patrick H., Thomas O. Binford, Boris Katz and Michael Lowry, "Learning Physical Descriptions from Functional Definitions, Examples, and Precedents," *AAAI-83*, 433-439, 1983.