Deriving Backtracking Monad Transformers

Functional Pearl

Ralf Hinze
Institut für Informatik III
Universität Bonn
Römerstraße 164, 53117 Bonn, Germany
ralf@informatik.uni-bonn.de

ABSTRACT

In a paper about pretty printing J. Hughes introduced two fundamental techniques for deriving programs from their specification, where a specification consists of a signature and properties that the operations of the signature are required to satisfy. Briefly, the first technique, the term implementation, represents the operations by terms and works by defining a mapping from operations to observations — this mapping can be seen as defining a simple interpreter. The second, the context-passing implementation, represents operations as functions from their calling context to observations. We apply both techniques to derive a backtracking monad transformer that adds backtracking to an arbitrary monad. In addition to the usual backtracking operations — failure and nondeterministic choice — the prolog cut and an operation for delimiting the effect of a cut are supported.

Categories and Subject Descriptors

D.1.1 [**Programming Techniques**]: Applicative (Functional) Programming; D.3.2 [**Programming Languages**]: Language Classifications—*applicative (functional) languages*; D.3.3 [**Programming Languages**]: Language Constructs and Features—*control structures; polymorphism*; F.3.1 [**Logics and Meanings of Programs**]: Specifying and Verifying and Reasoning about Programs—*specification techniques*; F.3.2 [**Logics and Meanings of Programs**]: Semantics of Programming Languages—*algebraic approaches to semantics*; F.3.3 [**Logics and Meanings of Programs**]: Studies of Program Constructs—*control primitives*

General Terms

Design, languages, theory, verification

Keywords

Program derivation, monads, monad transformers, backtracking, cut, continuations, Haskell, Prolog

ICFP '00, Montreal, Canada.

1. INTRODUCTION

Why should one derive a program from its specification? Ideally, a derivation explains and motivates the various design choices taken in a particular implementation. At best a derivation eliminates the need for so-called eureka steps, which are usually inevitable if a program is explained, say, by means of example.

In a paper about pretty printing J. Hughes [6] introduced two fundamental techniques for deriving programs from their specification. Both techniques provide the programmer with considerable guidance in the process of program derivation. To illustrate their utility and versatility we apply the framework to derive several monad transformers, which among other things add backtracking to an arbitrary monad.

Briefly, a monad transformer is a mapping on monads that augments a given monad by a certain computational feature such as state, exceptions, or nondeterminism. Traditionally, monad transformers are introduced in a single big eureka step. Even the recent introductory textbook on functional programming [2] fails to explain the particular definitions of monad transformers. After defining an exception monad transformer R. Bird remarks: "Why have we chosen to write [. . .]? The answer is: because it works.". Building upon Hughes' techniques we will try to provide a more satisfying answer. The reader should be prepared, however, that the results are somewhat different from the standard textbook examples.

The paper is organized as follows. Sec. 2 reviews monads and monad transformers. Sec. 3 introduces Hughes' techniques by means of a simple example. Sec. 4 applies the framework to derive a backtracking monad transformer that adds backtracking to an arbitrary monad. Finally, Sec. 5 extends the design of Sec. 4 to include additional control constructs: Prolog's cut and an operation for delimiting the effect of cut. Finally, Sec. 6 concludes and points out directions for future work.

2. PRELIMINARIES

Monads have been proposed by Moggi as a means to structure denotational semantics [11, 12]. Wadler popularized Moggi's idea in the functional programming community by using monads to structure functional programs [15, 16, 17]. In Haskell 98 [13] monads are captured by the class definition in Fig. 1. The essential idea of monads is to distinguish between *computations* and *values*. This distinction is reflected on the type level: an element of $m\ a$ represents a computation that yields a value of type a. The trivial computation

$$
\begin{array}{lll}
\multicolumn{3}{l}{\mathbf{class}\ \mathit{Monad}\ m\ \mathbf{where}} \\
\quad \mathit{return} & :: & a \to m\ a \\
\quad (\gg\!=) & :: & m\ a \to (a \to m\ b) \to m\ b \\
\quad (\gg) & :: & m\ a \to m\ b \to m\ b \\
\quad \mathit{fail} & :: & \mathit{String} \to m\ a \\
\quad m \gg n & = & m \gg\!= \mathit{const}\ n \\
\quad \mathit{fail}\ s & = & \mathit{error}\ s
\end{array}
$$

Figure 1: The *Monad* class.

that immediately returns the value a is denoted $\mathit{return}\ a$. The operator $(\gg\!=)$, commonly called 'bind', combines two computations: $m \gg\!= k$ applies k to the result of the computation m. The derived operation $(\gg)$ provides a handy shortcut if one is not interested in the result of the first computation. The operation *fail* is useful for signaling error conditions and will be used to this effect. Note that *fail* does not stem from the mathematical concept of a monad, but has been added to the monad class for pragmatic reasons, see [13, Sec. 3.14].

The operations are required to satisfy the following so-called *monad laws*.

$$
\begin{array}{rcll}
\mathit{return}\ a \gg\!= k & = & k\ a & \text{(M1)} \\
m \gg\!= \mathit{return} & = & m & \text{(M2)} \\
(m \gg\!= k_1) \gg\!= k_2 & = & m \gg\!= (\lambda a \to k_1\ a \gg\!= k_2) & \text{(M3)}
\end{array}
$$

For an explanation of the laws we refer the reader to [2, Sec. 10.3]. Note that *fail* is intentionally left unspecified.

Different monads are distinguished by the computational features they support. Each computational feature is typically accessed through a number of additional operations. For instance, a backtracking monad additionally supports the operations *false* and $(\shortmid)$ denoting failure and nondeterministic choice. It is relatively easy to construct a monad that supports only a single computational feature. Unfortunately, there is no uniform way of combining two monads, which support different computational features. The reason is simply that two features may interact in different ways. There is, however, a uniform method for augmenting a given monad by a certain computational feature. This method is captured by the following class definition which introduces *monad transformers* [9].

$$
\begin{array}{lll}
\multicolumn{3}{l}{\mathbf{class}\ \mathit{Transformer}\ \tau\ \mathbf{where}} \\
\quad \mathit{promote} & :: & (\mathit{Monad}\ m) \Rightarrow m\ a \to \tau\ m\ a \\
\quad \mathit{observe} & :: & (\mathit{Monad}\ m) \Rightarrow \tau\ m\ a \to m\ a
\end{array}
$$

A monad transformer is basically a type constructor τ that takes a monad m to a monad $\tau\ m$. It must additionally provide two operations: an operation for embedding computations from the underlying monad into the transformed monad and an inverse operation, which allows us to observe 'augmented' computations in the underlying monad. Since *observe* forgets structure, it will in general be a partial function. In what follows we will abbreviate *observe* by ω and *promote* by π. Turning to the laws we require promotion to respect the monad operations.

$$
\begin{array}{rcll}
\pi\ (\mathit{return}\ a) & = & \mathit{return}\ a & \text{(P1)} \\
\pi\ (m \gg\!= k) & = & \pi\ m \gg\!= (\pi \cdot k) & \text{(P2)}
\end{array}
$$

These laws determine π as a *monad morphism*. In general, π should respect every operation the underlying monad provides in order to guarantee that a program that does not use new features behaves the same in the underlying and in the transformed monad. The counterpart of π is not quite a monad morphism.

$$
\begin{array}{rcll}
\omega\ (\mathit{return}\ a) & = & \mathit{return}\ a & \text{(O1)} \\
\omega\ (\pi\ m \gg\!= k) & = & m \gg\!= (\omega \cdot k) & \text{(O2)}
\end{array}
$$

The second law is weaker than the corresponding law for π. It is unreasonable to expect more since computations in $\tau\ m$ can, in general, not be mimicked in m.

3. ADDING ABNORMAL TERMINATION

This section reviews Hughes' technique by means of a simple example. We show how to augment a given monad by an operation that allows one to terminate a computation abnormally. Monads with additional features are introduced as subclasses of *Monad*.

$$
\begin{array}{lll}
\mathbf{type}\ \mathit{Exception} & = & \mathit{String} \\
\multicolumn{3}{l}{\mathbf{class}\ (\mathit{Monad}\ m) \Rightarrow \mathit{Raise}\ m\ \mathbf{where}} \\
\quad \mathit{raise} & :: & \mathit{Exception} \to m\ a
\end{array}
$$

The call $\mathit{raise}\ e$ terminates the current computation. This property is captured by the law:

$$
\mathit{raise}\ e \gg\!= k \quad = \quad \mathit{raise}\ e, \qquad \text{(R1)}
$$

which formalizes that $\mathit{raise}\ e$ is a left zero of $(\gg\!=)$. Now, let us try to derive a monad transformer for this feature. Beforehand, we must determine how $\mathit{raise}\ e$ is observed in the base monad. We specify:

$$
\omega\ (\mathit{raise}\ e) \quad = \quad \mathit{fail}\ e, \qquad \text{(O3)}
$$

which appears to be the only reasonable choice since we know nothing of the underlying monad.

Remark. We do not consider an operation for trapping exceptions (such as *handle*) in order to keep the introductory example short and simple. It is worth noting, however, that the derivation of a fully-fledged exception monad transformer proceeds similar to the derivation given in Sec. 5.

3.1 A term implementation

The term implementation represents operations simply by terms of the algebra and works by defining an interpreter for the language. Since we have four operations — *return*, $(\gg\!=)$, *raise*, and π — the datatype that implements the term algebra consequently comprises four constructors. We adopt the convention that monad transformers are given names that are all in upper case. For the constructor names we re-use the names of the operations with the first letter in upper case; operators like $(\gg\!=)$ are prefixed by a colon.

$$
\begin{array}{ll}
\multicolumn{2}{l}{\mathbf{data}\ \mathit{RAISE}\ m\ a} \\
\quad = & \mathit{Return}\ a \\
\quad \mid & \forall b.\ (\mathit{RAISE}\ m\ b) :\gg\!= (b \to \mathit{RAISE}\ m\ a) \\
\quad \mid & \mathit{Raise}\ \mathit{Exception} \\
\quad \mid & \mathit{Promote}\ (m\ a)
\end{array}
$$

Note that the definition involves an existentially quantified type variable [8] in the type of $(:\gg\!=)$. We use GHC/Hugs syntax for existential quantification: the existentially quantified variable is bound by an explicit *universal* quantifier written *before* the constructor.

$$\begin{array}{lcl}
\mathbf{data}\ \mathit{RAISE}\ m\ a & = & \mathit{Return}\ a \\
 & \mid & \forall b.(\mathit{RAISE}\ m\ b) :\gg\!= (b \to \mathit{RAISE}\ m\ a) \\
 & \mid & \mathit{Raise}\ \mathit{Exception} \\
 & \mid & \mathit{Promote}\ (m\ a) \\
\multicolumn{3}{l}{\mathbf{instance}\ \mathit{Monad}\ (\mathit{RAISE}\ m)\ \mathbf{where}} \\
\quad \mathit{return} & = & \mathit{Return} \\
\quad (\gg\!=) & = & (:\gg\!=) \\
\multicolumn{3}{l}{\mathbf{instance}\ \mathit{Raise}\ (\mathit{RAISE}\ m)\ \mathbf{where}} \\
\quad \mathit{raise} & = & \mathit{Raise} \\
\multicolumn{3}{l}{\mathbf{instance}\ \mathit{Transformer}\ \mathit{RAISE}\ \mathbf{where}} \\
\quad \mathit{promote} & = & \mathit{Promote} \\
\quad \mathit{observe}\ (\mathit{Return}\ a) & = & \mathit{return}\ a \\
\quad \mathit{observe}\ (\mathit{Return}\ a :\gg\!= k) & = & \mathit{observe}\ (k\ a) \\
\quad \mathit{observe}\ ((m :\gg\!= k_1) :\gg\!= k_2) & = & \mathit{observe}\ (m :\gg\!= (\lambda a \to k_1\ a :\gg\!= k_2)) \\
\quad \mathit{observe}\ (\mathit{Raise}\ e :\gg\!= k) & = & \mathit{fail}\ e \\
\quad \mathit{observe}\ (\mathit{Promote}\ m :\gg\!= k) & = & m \gg\!= (\mathit{observe} \cdot k) \\
\quad \mathit{observe}\ (\mathit{Raise}\ e) & = & \mathit{fail}\ e \\
\quad \mathit{observe}\ (\mathit{Promote}\ m) & = & m
\end{array}$$

Figure 2: A term implementation of *RAISE*.

Now, each of the operations *return*, $(\gg\!=)$, *raise*, and π is implemented by the corresponding constructor. In other words, the operations do nothing. All the work is performed by ω which can be seen as defining a tiny interpreter for the monadic language. Except for one case the definition of ω is straightforward.

$$\begin{array}{lcl}
\omega\ (\mathit{Return}\ a) & = & \mathit{return}\ a \\
\omega\ (m :\gg\!= k) & = & \\
\omega\ (\mathit{Raise}\ e) & = & \mathit{fail}\ e \\
\omega\ (\mathit{Promote}\ m) & = & m
\end{array}$$

Can we fill in the blank on the right-hand side? It appears impossible to define $\omega\ (m :\gg\!= k)$ in terms of its constituents. The only way out of this dilemma is to make a further case distinction on m:

$$\begin{array}{lcl}
\omega\ (\mathit{Return}\ a :\gg\!= k) & = & \omega\ (k\ a) \\
\omega\ ((m :\gg\!= k_1) :\gg\!= k_2) & = & \omega\ (m :\gg\!= (\lambda a \to k_1\ a :\gg\!= k_2)) \\
\omega\ (\mathit{Raise}\ e :\gg\!= k) & = & \mathit{fail}\ e \\
\omega\ (\mathit{Promote}\ m :\gg\!= k) & = & m \gg\!= (\omega \cdot k).
\end{array}$$

Voilà. Each equation is a simple consequence of the monad laws and the laws for ω. In particular, the second equation employs (M3), the associative law for $(\gg\!=)$, to reduce the size of $(:\gg\!=)$'s first argument. This rewrite step is analogous to rotating a binary tree to the right. Fig. 2 summarizes the term implementation. Note that in the sequel we will omit trivial instance declarations like *Monad* (*RAISE* m) and *Raise* (*RAISE* m).

What about correctness? First of all, the definition of ω is exhaustive. It is furthermore terminating since the size of $(:\gg\!=)$'s left argument is steadily decreasing. We can establish termination using a so-called *polynomial interpretation* of the operations [4]:

$$\begin{array}{rcl@{\qquad}rcl}
\mathit{Return}_\tau\ a & = & 1 & \mathit{Raise}_\tau\ e & = & 1 \\
m :\gg_\tau n & = & 2 \times m + n & \mathit{Promote}_\tau\ m & = & 1.
\end{array}$$

A multivariate polynomial op_τ of n variables is associated with each n-ary operation op. For each equation $\omega\ \ell = \dots\ \omega\ r\ \dots$ we must show that $\tau\ \ell > \tau\ r$ for all variables (ranging over positive integers) where τ is given by $\tau(op\ e_1 \dots e_n) = op_\tau(\tau\ e_1)\dots(\tau\ e_n)$. Note that we consider bind only for the special case that the result of the first argument is ignored. The inclusion of $m :\gg\!= k$ in its full generality is feasible but technically more involving since the interpretation of k depends on the value m computes.

Does the implementation satisfy its specification? Since we are working in the free algebra, the laws do not hold: the expressions *Return* a and *Return* $a :\gg\!= \mathit{Return}$, for example, are distinct, unrelated terms. The laws of the specification only hold *under observation*. The monad laws become:

$$\begin{array}{rcl}
\omega\ (\mathit{return}\ a \gg\!= k) & = & \omega\ (k\ a) \\
\omega\ (m \gg\!= \mathit{return}) & = & \omega\ m \\
\omega\ ((m \gg\!= k_1) \gg\!= k_2) & = & \omega\ (m \gg\!= (\lambda a \to k_1\ a \gg\!= k_2)).
\end{array}$$

The first and the third are direct consequences of ω's definition. The second can be shown by induction on m. Fortunately, we can live with the weakened laws, since the only way to run computations of type *RAISE* m is to use ω.

3.2 A simplified term implementation

Can we do better than the naive term implementation? A major criticism of the first attempt is that the operations do not exploit the algebraic laws. It is conceivable that we can work with a subset of the term algebra. For instance, we need not represent both *Raise* e and *Raise* $e :\gg\!= \mathit{Return}$. A rather systematic way to determine the required subset of terms is to program a simplifier for the datatype *RAISE*, which exploits the algebraic laws as far as possible. It turns out that we only need to modify ω slightly.

$$\begin{array}{lcl}
\sigma & :: & \mathit{RAISE}\ m\ a \to \mathit{RAISE}\ m\ a \\
\sigma\ (\mathit{Return}\ a) & = & \mathit{Return}\ a \\
\sigma\ (\mathit{Return}\ a :\gg\!= k) & = & \sigma\ (k\ a) \\
\sigma\ ((m :\gg\!= k_1) :\gg\!= k_2) & = & \sigma\ (m :\gg\!= (\lambda a \to k_1\ a :\gg\!= k_2)) \\
\sigma\ (\mathit{Raise}\ e :\gg\!= k) & = & \mathit{Raise}\ e \\
\sigma\ (\mathit{Promote}\ m :\gg\!= k) & = & \mathit{Promote}\ m :\gg\!= (\sigma \cdot k) \\
\sigma\ (\mathit{Raise}\ e) & = & \mathit{Raise}\ e
\end{array}$$

$$\sigma\ (\mathit{Promote}\ m) \quad = \quad \mathit{Promote}\ m$$

Inspecting the right hand sides we see that we require $(:\gg=)$ only in conjunction with *Promote*. Since $\pi\ m$ is furthermore equivalent to $\pi\ m \gg= \mathit{return}$ we can, in fact, restrict ourselves to the following subset of the term algebra.

$$\begin{array}{lll}
\mathbf{data}\ \mathit{RAISE}\ m\ a & & \\
\quad = & \mathit{Return}\ a & \\
\quad \mid & \forall b.\ \mathit{PromoteBind}\ (m\ b)\ (b \to \mathit{RAISE}\ m\ a) & \\
\quad \mid & \mathit{Raise}\ \mathit{Exception} &
\end{array}$$

Following Hughes [6] we call elements of the new datatype *simplified terms*. We avoid the term normal form or canonical form since distinct terms may not necessarily be semantically different. For instance, *return a* can be represented both by *Return a* and *PromoteBind (return a) Return*. Nonetheless, using this representation the definition of ω is much simpler. It is, in fact, directly based on the laws (O1)–(O3). The complete implementation appears in Fig. 3. If we are only interested in defining a monad (not a monad transformer), then we can omit the constructor *PromoteBind*. The resulting datatype corresponds exactly to the standard definition of the exception monad.

What about efficiency? The naive implementation — or rather, the first definition of ω has a running time that is proportional to the size of the computation. Unfortunately, the 'improved' term implementation has a quadratic worst-case behaviour. Consider the expression

$$\omega\ (\cdots\ ((\pi\ (\mathit{return}\ 0) \gg= \mathit{inc}) \gg= \mathit{inc})\ \cdots \gg= \mathit{inc}).$$

where *inc* is given by $\mathit{inc}\ n = \pi\ (\mathit{return}\ (n+1))$. Since the amortized running time of bind is proportional to the size of its first argument, it takes $O(n^2)$ steps to evaluate the expression above. The situation is analogous to flattening a binary tree. Bad luck.

3.3 A context-passing implementation

Since we cannot improve the implementation of the operations without sacrificing the runtime efficiency, let us try to improve the definition of ω. While rewriting ω we will work out a specification for the final *context-passing implementation*. For a start, we can avoid some pattern matching if we specialize ω for $\mathit{op} \gg= k$. To this end we replace the equations concerning $(\gg=)$ by the single equation

$$\omega\ (\mathit{op} :\gg= c) \quad = \quad \omega_1\ \mathit{op}\ c$$

and define ω_1 by

$$\begin{array}{lcl}
\omega_1\ (\mathit{Return}\ a)\ c & = & \omega\ (c\ a) \\
\omega_1\ (m :\gg= k)\ c & = & \omega_1\ m\ (\lambda a \to k\ a :\gg= c) \\
\omega_1\ (\mathit{Raise}\ e)\ c & = & \mathit{fail}\ e \\
\omega_1\ (\mathit{Promote}\ m)\ c & = & m \gg= \lambda a \to \omega\ (c\ a).
\end{array}$$

Interestingly, the parameter c is used twice in conjunction with ω. In an attempt to eliminate the mutual recursive dependence on ω we could try to pass $\omega \cdot c$ as a parameter instead of c. This variation of ω_1, which we call $\underline{\omega}$, can be specified as follows.

$$\begin{array}{ll}
& \underline{\omega}\ \mathit{op}\ \underline{c} = \omega\ (\mathit{op} :\gg= c) \\
\Longleftarrow & \forall a.\ \underline{c}\ a = \omega\ (c\ a)
\end{array} \qquad (1)$$

Let us derive the definition of $\underline{\omega}$ for $\mathit{op} = \mathit{Return}\ a$. We assume that precondition (1) holds — note that the equation number refers to the precondition only — and reason:

$$\begin{array}{ll}
& \underline{\omega}\ (\mathit{Return}\ a)\ \underline{c} \\
= & \quad \{\ \text{specification and assumption (1)}\ \} \\
& \omega\ (\mathit{Return}\ a :\gg= c) \\
= & \quad \{\ \text{definition } \omega\ \} \\
& \omega\ (c\ a) \\
= & \quad \{\ \text{assumption (1)}\ \} \\
& \underline{c}\ a.
\end{array}$$

The calculations for *Promote m* and *Raise e* are similar. It remains to infer the definition for $\mathit{op} = (m :\gg= k)$:

$$\begin{array}{ll}
& \underline{\omega}\ (m :\gg= k)\ \underline{c} \\
= & \quad \{\ \text{specification and assumption (1)}\ \} \\
& \omega\ ((m :\gg= k) :\gg= c) \\
= & \quad \{\ \text{definition } \omega\ \} \\
& \omega\ (m :\gg= (\lambda a \to k\ a :\gg= c)) \\
= & \quad \{\ \text{specification}\ \} \\
& \underline{\omega}\ m\ (\lambda a \to \omega\ (k\ a :\gg= c)) \\
= & \quad \{\ \text{specification and assumption (1)}\ \} \\
& \underline{\omega}\ m\ (\lambda a \to \underline{\omega}\ (k\ a)\ \underline{c}).
\end{array}$$

Voilà. The dependence on ω has vanished. To summarize, $\underline{\omega}$ is given by

$$\begin{array}{lcl}
\underline{\omega}\ (\mathit{Return}\ a) & = & \lambda\underline{c} \to \underline{c}\ a \\
\underline{\omega}\ (m :\gg= k) & = & \lambda\underline{c} \to \underline{\omega}\ m\ (\lambda a \to \underline{\omega}\ (k\ a)\ \underline{c}) \\
\underline{\omega}\ (\mathit{Raise}\ e) & = & \lambda\underline{c} \to \mathit{fail}\ e \\
\underline{\omega}\ (\mathit{Promote}\ m) & = & \lambda\underline{c} \to m \gg= \underline{c}.
\end{array}$$

Note that the constructors appear only on the left-hand sides. This means that we are even able to remove the interpretative layer, ie *return a* can be implemented directly by $\lambda\underline{c} \to \underline{c}\ a$ instead of *Return*. In general, we consistently replace $\underline{\omega}\ \mathit{op}$ by *op*. Silently, we have converted the term implementation into a *context-passing implementation*. To see why the term 'context-passing' is appropriate, consider the final specification of the context-passing implementation.

$$\begin{array}{ll}
& \mathit{op}\ \underline{c} = \omega\ (\mathit{op} \gg= c) \\
\Longleftarrow & \forall a.\ \underline{c}\ a = \omega\ (c\ a)
\end{array} \qquad (2)$$

The parameter $\underline{c}$ of *op* can be seen as a representation of *op*'s calling context $\omega\ (\bullet \gg= c)$ — we represent a context by an expression that has a hole in it. This is the nub of the story: every operation knows the context in which it is called and it is furthermore able to access and to rearrange the context. This gives the implementor a much greater freedom of manoeuvre as compared to the simplified term algebra. For instance, $(\gg=)$ can use the associative law to improve efficiency. By contrast, $(\gg=)$ of the simplified term variety does not know of any outer binds and consequently falls into the efficiency trap.

It is quite instructive to infer the operations of the context-passing implementation from scratch using the specification above. Fig. 4 summarizes the calculations. Interestingly, each monad law, the law for *raise*, and each law for ω is invoked exactly once. In other words, the laws of the specification are necessary and sufficient for deriving an implementation.

It remains to determine the type of the new monad transformer. This is most easily accomplished by inspecting the

$$\begin{array}{lcl}
\textbf{data}\ \mathit{RAISE}\ m\ a & = & \mathit{Return}\ a \\
 & | & \forall b.\mathit{PromoteBind}\ (m\ b)\ (b \to \mathit{RAISE}\ m\ a) \\
 & | & \mathit{Raise}\ \mathit{Exception} \\
\textbf{instance}\ \mathit{Monad}\ (\mathit{RAISE}\ m)\ \textbf{where} & & \\
\quad \mathit{return}\ a & = & \mathit{Return}\ a \\
\quad \mathit{Return}\ a \gg\!= k & = & k\ a \\
\quad (\mathit{PromoteBind}\ m\ k_1) \gg\!= k_2 & = & \mathit{PromoteBind}\ m\ (\lambda a \to k_1\ a \gg\!= k_2) \\
\quad \mathit{Raise}\ e \gg\!= k & = & \mathit{Raise}\ e \\
\textbf{instance}\ \mathit{Raise}\ (\mathit{RAISE}\ m)\ \textbf{where} & & \\
\quad \mathit{raise}\ e & = & \mathit{Raise}\ e \\
\textbf{instance}\ \mathit{Transformer}\ \mathit{RAISE}\ \textbf{where} & & \\
\quad \mathit{promote}\ m & = & \mathit{PromoteBind}\ m\ \mathit{Return} \\
\quad \mathit{observe}\ (\mathit{Return}\ a) & = & \mathit{return}\ a \\
\quad \mathit{observe}\ (\mathit{PromoteBind}\ m\ k) & = & m \gg\!= (\mathit{observe} \cdot k) \\
\quad \mathit{observe}\ (\mathit{Raise}\ e) & = & \mathit{fail}\ e
\end{array}$$

Figure 3: A simplified term implementation of *RAISE*.

definition of π. Note that $\pi\ m$ equals $(\gg\!=)\ m$ and recall that $(\gg\!=)$ possesses the type $\forall a.\forall b.m\ a \to (a \to m\ b) \to m\ b$ which is equivalent to $\forall a.m\ a \to (\forall b.(a \to m\ b) \to m\ b)$. Consequently, the new transformer has type $\forall b.(a \to m\ b) \to m\ b$. So, while the term implementation requires existential quantification, the context-passing implementation makes use of universal quantification. The final implementation appears in Fig. 5.[1] The cognoscenti would certainly recognize that the implementation is identical with the definition of the *continuation monad transformer* [9]. Only the types are different: *RAISE* involves rank-2 types while the continuation monad transformer is additionally parameterized with the answer type: $\mathit{CONT}\ \mathit{ans}\ m\ a = (a \to m\ \mathit{ans}) \to m\ \mathit{ans}$. The transformer $\mathit{RAISE}\ m$ constitutes the smallest extension of m that allows one to add *raise*. Note, for instance, that *callcc* is definable in $\mathit{CONT}\ \mathit{ans}\ m$ but not in $\mathit{RAISE}\ m$. We will see in Sec. 4.3 that rank-2 types have advantages over parameterized types.

4. ADDING BACKTRACKING

By definition, a *backtracking monad* is a monad with two additional operations: the constant *false*, which denotes failure, and the binary operation $(\shortmid)$, which denotes nondeterministic choice. The class definition contains a third operation, termed *cons*, which provides a handy shortcut for $\mathit{return}\ a \shortmid m$.

$$\begin{array}{lcl}
\multicolumn{3}{l}{\textbf{class}\ (\mathit{Monad}\ m) \Rightarrow \mathit{Backtr}\ m\ \textbf{where}} \\
\quad \mathit{false} & :: & m\ a \\
\quad (\shortmid) & :: & m\ a \to m\ a \to m\ a \\
\quad \mathit{cons} & :: & a \to m\ a \to m\ a \\
\quad \mathit{cons}\ a\ m & = & \mathit{return}\ a \shortmid m
\end{array}$$

The operations are required to satisfy the following laws.

$$\mathit{false} \shortmid m = m \tag{B1}$$

$$m \shortmid \mathit{false} = m \tag{B2}$$

$$(m \shortmid n) \shortmid o = m \shortmid (n \shortmid o) \tag{B3}$$

$$\mathit{false} \gg\!= k = \mathit{false} \tag{B4}$$

$$(m \shortmid n) \gg\!= k = (m \gg\!= k) \shortmid (n \gg\!= k) \tag{B5}$$

That is, *false* and $(\shortmid)$ form a monoid; *false* is a left zero of $(\gg\!=)$, and $(\gg\!=)$ distributes leftward through $(\shortmid)$. Now, since we aim at defining a backtracking monad transformer, we must also specify the interaction of promoted operations with $(\shortmid)$:

$$(\pi\ m \gg\!= k) \shortmid n = \pi\ m \gg\!= \lambda a \to k\ a \shortmid n. \tag{B6}$$

Consider $\pi\ m$ as a deterministic computation, ie a computation that succeeds exactly once. Then (B6) formalizes our intuition that a deterministic computation can be pushed out of a disjunction's left branch. Finally, we must specify how the backtracking operations are observed in the base monad.

$$\omega\ \mathit{false} = \mathit{fail}\ \texttt{"false"} \tag{O4}$$

$$\omega\ (\mathit{return}\ a \shortmid m) = \mathit{return}\ a \tag{O5}$$

So we can observe the first answer of a nondeterministic computation.

4.1 A term implementation

The free term algebra of the backtracking monad is given by the following type definition.

$$\begin{array}{cl}
\multicolumn{2}{l}{\textbf{data}\ \mathit{BACKTR}\ m\ a} \\
= & \mathit{Return}\ a \\
| & \forall b.\ (\mathit{BACKTR}\ m\ b) :\gg\!= (b \to \mathit{BACKTR}\ m\ a) \\
| & \mathit{False} \\
| & \mathit{BACKTR}\ m\ a :\shortmid \mathit{BACKTR}\ m\ a \\
| & \mathit{Promote}\ (m\ a)
\end{array}$$

Let us try to derive an interpreter for this language. The definition of the base cases follows immediately from the specification. For $m :\gg\!= k$ we obtain:

$$\begin{array}{lcl}
\omega\ (\mathit{Return}\ a :\gg\!= k) & = & \omega\ (k\ a) \\
\omega\ ((m :\gg\!= k_1) :\gg\!= k_2) & = & \omega\ (m :\gg\!= (\lambda a \to k_1\ a :\gg\!= k_2)) \\
\omega\ (\mathit{False} :\gg\!= k) & = & \mathit{fail}\ \texttt{"false"} \\
\omega\ ((m :\shortmid n) :\gg\!= k) & = & \omega\ ((m :\gg\!= k) :\shortmid (n :\gg\!= k)) \\
\omega\ (\mathit{Promote}\ m :\gg\!= k) & = & m \gg\!= (\omega \cdot k).
\end{array}$$

[1] Note that *RAISE* must actually be defined using **newtype** instead of **type**. This, however, introduces an additional data constructor that affects the readability of the code. Instead we employ **type** declarations as if they worked as **newtype** declarations.

$$
\begin{array}{ll}
 & (\mathit{return}\ a)\ \underline{c} \\
= & \quad \{\ \text{specification and assumption (2)}\ \} \\
 & \mathit{observe}\ (\mathit{return}\ a \gg\!= c) \\
= & \quad \{\ \text{(M1)}\ \} \\
 & \mathit{observe}\ (c\ a) \\
= & \quad \{\ \text{assumption (2)}\ \} \\
 & \underline{c}\ a
\end{array}
$$

$$
\begin{array}{ll}
 & (m \gg\!= k)\ \underline{c} \\
= & \quad \{\ \text{specification and assumption (2)}\ \} \\
 & \mathit{observe}\ ((m \gg\!= k) \gg\!= c) \\
= & \quad \{\ \text{(M3)}\ \} \\
 & \mathit{observe}\ (m \gg\!= (\lambda a \to k\ a \gg\!= c)) \\
= & \quad \{\ \text{specification}\ \} \\
 & m\ (\lambda a \to \mathit{observe}\ (k\ a \gg\!= c)) \\
= & \quad \{\ \text{specification and assumption (2)}\ \} \\
 & m\ (\lambda a \to k\ a\ \underline{c})
\end{array}
$$

$$
\begin{array}{ll}
 & (\mathit{raise}\ e)\ \underline{c} \\
= & \quad \{\ \text{specification and assumption (2)}\ \} \\
 & \mathit{observe}\ (\mathit{raise}\ e \gg\!= c) \\
= & \quad \{\ \text{(R1)}\ \} \\
 & \mathit{observe}\ (\mathit{raise}\ e) \\
= & \quad \{\ \text{(O3)}\ \} \\
 & \mathit{fail}\ e
\end{array}
$$

$$
\begin{array}{ll}
 & (\mathit{promote}\ m)\ \underline{c} \\
= & \quad \{\ \text{specification and assumption (2)}\ \} \\
 & \mathit{observe}\ (\mathit{promote}\ m \gg\!= c) \\
= & \quad \{\ \text{(O2)}\ \} \\
 & m \gg\!= \lambda a \to \mathit{observe}\ (c\ a) \\
= & \quad \{\ \text{assumption (2)}\ \} \\
 & m \gg\!= \underline{c}
\end{array}
$$

$$
\begin{array}{ll}
 & \mathit{observe}\ m \\
= & \quad \{\ \text{(M2)}\ \} \\
 & \mathit{observe}\ (m \gg\!= \mathit{return}) \\
= & \quad \{\ \text{specification}\ \} \\
 & m\ (\lambda a \to \mathit{observe}\ (\mathit{return}\ a)) \\
= & \quad \{\ \text{(O1)}\ \} \\
 & m\ \mathit{return}
\end{array}
$$

Figure 4: Deriving a context-passing implementation of *RAISE*.

$$
\begin{array}{lll}
\textbf{type}\ \mathit{RAISE}\ m\ a & = & \forall b.(a \to m\ b) \to m\ b \\
\textbf{instance}\ (\mathit{Monad}\ m) \Rightarrow \mathit{Monad}\ (\mathit{RAISE}\ m)\ \textbf{where} & & \\
\quad \mathit{return}\ a & = & \lambda\underline{c} \to \underline{c}\ a \\
\quad m \gg\!= k & = & \lambda\underline{c} \to m\ (\lambda a \to k\ a\ \underline{c}) \\
\textbf{instance}\ (\mathit{Monad}\ m) \Rightarrow \mathit{Raise}\ (\mathit{RAISE}\ m)\ \textbf{where} & & \\
\quad \mathit{raise}\ e & = & \lambda\underline{c} \to \mathit{fail}\ e \\
\textbf{instance}\ \mathit{Transformer}\ \mathit{RAISE}\ \textbf{where} & & \\
\quad \mathit{promote}\ m & = & \lambda\underline{c} \to m \gg\!= \underline{c} \\
\quad \mathit{observe}\ m & = & m\ \mathit{return}
\end{array}
$$

Figure 5: A context-passing implementation of *RAISE*.

Similarly, for $m \mathbin{:\!|} n$ we make a case distinction on m:

$$
\begin{array}{lll}
\omega\ (\mathit{Return}\ a \mathbin{:\!|} \mathrm{f}) & = & \mathit{return}\ a \\
\omega\ (m \mathbin{:\!\gg\!=} k \mathbin{:\!|} \mathrm{f}) & = & \\
\omega\ (\mathit{False} \mathbin{:\!|} \mathrm{f}) & = & \omega\ \mathrm{f} \\
\omega\ ((m \mathbin{:\!|} n) \mathbin{:\!|} \mathrm{f}) & = & \omega\ (m \mathbin{:\!|} (n \mathbin{:\!|} \mathrm{f})) \\
\omega\ (\mathit{Promote}\ m \mathbin{:\!|} \mathrm{f}) & = & m.
\end{array}
$$

Unfortunately, one case remains. There is no obvious way to simplify $\omega\ (m \mathbin{:\!\gg\!=} k \mathbin{:\!|} \mathrm{f})$. As usual, we help ourselves by making a further case distinction on m.

$$
\begin{array}{lll}
\omega\ ((\mathit{Return}\ a \mathbin{:\!\gg\!=} k) \mathbin{:\!|} \mathrm{f}) & = & \omega\ (k\ a \mathbin{:\!|} \mathrm{f}) \\
\omega\ (((m \mathbin{:\!\gg\!=} k_1) \mathbin{:\!\gg\!=} k_2) \mathbin{:\!|} \mathrm{f}) & = & \omega\ ((m \mathbin{:\!\gg\!=} (\lambda a \to k_1\ a \\
 & & \qquad\qquad \mathbin{:\!\gg\!=} k_2)) \mathbin{:\!|} \mathrm{f}) \\
\omega\ ((\mathit{False} \mathbin{:\!\gg\!=} k) \mathbin{:\!|} \mathrm{f}) & = & \omega\ \mathrm{f} \\
\omega\ (((m \mathbin{:\!|} n) \mathbin{:\!\gg\!=} k) \mathbin{:\!|} \mathrm{f}) & = & \omega\ ((m \mathbin{:\!\gg\!=} k) \\
 & & \qquad\qquad \mathbin{:\!|} ((n \mathbin{:\!\gg\!=} k) \mathbin{:\!|} \mathrm{f})) \\
\omega\ ((\mathit{Promote}\ m \mathbin{:\!\gg\!=} k) \mathbin{:\!|} \mathrm{f}) & = & m \gg\!= \lambda a \to \omega\ (k\ a \mathbin{:\!|} \mathrm{f})
\end{array}
$$

Voilà. We have succeeded in building an interpreter for backtracking. Fig. 6 lists the complete implementation.

Now, what about correctness? Clearly, the case distinction is exhaustive. To establish termination we can use the following polynomial interpretation.

$$
\begin{array}{rclrcl}
\mathit{Return}_{\tau}\ a & = & 2 & m \mathbin{:\!|}_{\tau} n & = & 2 \times m + n \\
m \mathbin{:\!\gg}_{\tau} n & = & m^2 \times n & \mathit{Promote}_{\tau}\ m & = & 2 \\
\mathit{False}_{\tau} & = & 2 & & &
\end{array}
$$

As before, the laws of the specification only hold under observation.

4.2 A simplified term implementation

Let us take a brief look at the simplified term implementation. Inspecting the definition of ω — recall that a simplifier is likely to make the same case distinction as ω — we see that we need at most six terms: *False*, $\mathit{Return}\ a$, $\mathit{Return}\ a \mathbin{:\!|} \mathrm{f}$, $\mathit{Promote}\ m$, $\mathit{Promote}\ m \mathbin{:\!\gg\!=} k$, and $\mathit{Promote}\ m \mathbin{:\!|} \mathrm{f}$. We can eliminate three of them using $\mathit{return}\ a = \mathit{cons}\ a\ \mathit{false}$, $\pi\ m = \pi\ m \gg\!= \mathit{return}$, and $\pi\ m \mathbin{:\!|} \mathrm{f} = \pi\ m \gg\!= \lambda a \to \mathit{cons}\ a\ \mathrm{f}$. This explains the following definition of simplified terms.

$$
\begin{array}{ll}
\multicolumn{2}{l}{\textbf{data}\ \mathit{BACKTR}\ m\ a} \\
\quad = & \mathit{False} \\
\quad \mid & \mathit{Cons}\ a\ (\mathit{BACKTR}\ m\ a) \\
\quad \mid & \forall b.\ \mathit{PromoteBind}\ (m\ b)\ (b \to \mathit{BACKTR}\ m\ a)
\end{array}
$$

In essence, the simplified term algebra is an extension of the datatype of parametric lists with *False* corresponding to [] and *Cons* corresponding to (:). The additional constructor *PromoteBind* makes the difference between a monad and a monad transformer. Note that the standard list monad transformer, $\mathit{LIST}\ m\ a = m\ [a]$, can only be applied to so-called *commutative monads* [7]. By contrast, *BACKTR* works for arbitrary monads.

4.3 A context-passing implementation

In Sec. 3.3 we have seen that the context-passing implementation essentially removes the interpretative layer from the 'naive' term implementation. If we apply the same steps, we can derive very systematically a context-passing implementation of backtracking. We leave the details to the reader and sketch only the main points. First, from the case analysis ω performs we may conclude that the most complex context has the form $\omega\ (\bullet \gg\!= c \mid \mathrm{f})$. All other contexts can be rewritten into this form. Second, if we inspect the

$$
\begin{array}{lcl}
\textbf{data}\ \mathit{BACKTR}\ m\ a & = & \mathit{Return}\ a\\
 & \mid & \forall b.(\mathit{BACKTR}\ m\ b)\ {:}{\gg\!=}\ (b \to \mathit{BACKTR}\ m\ a)\\
 & \mid & \mathit{False}\\
 & \mid & \mathit{BACKTR}\ m\ a\ {:}{\mid}\ \mathit{BACKTR}\ m\ a\\
 & \mid & \mathit{Promote}\ (m\ a)\\
\textbf{instance}\ \mathit{Transformer}\ \mathit{BACKTR}\ \textbf{where} & & \\
\quad \mathit{promote} & = & \mathit{Promote}\\
\quad \mathit{observe}\ (\mathit{Return}\ a) & = & \mathit{return}\ a\\
\quad \mathit{observe}\ (\mathit{Return}\ a\ {:}{\gg\!=}\ k) & = & \mathit{observe}\ (k\ a)\\
\quad \mathit{observe}\ ((m\ {:}{\gg\!=}\ k_1)\ {:}{\gg\!=}\ k_2) & = & \mathit{observe}\ (m\ {:}{\gg\!=}\ (\lambda a \to k_1\ a\ {:}{\gg\!=}\ k_2))\\
\quad \mathit{observe}\ (\mathit{False}\ {:}{\gg\!=}\ k) & = & \mathit{fail}\ \texttt{"false"}\\
\quad \mathit{observe}\ ((m\ {:}{\mid}\ n)\ {:}{\gg\!=}\ k) & = & \mathit{observe}\ ((m\ {:}{\gg\!=}\ k)\ {:}{\mid}\ (n\ {:}{\gg\!=}\ k))\\
\quad \mathit{observe}\ (\mathit{Promote}\ m\ {:}{\gg\!=}\ k) & = & m \gg\!= (\mathit{observe} \cdot k)\\
\quad \mathit{observe}\ \mathit{False} & = & \mathit{fail}\ \texttt{"false"}\\
\quad \mathit{observe}\ (\mathit{Return}\ a\ {:}{\mid}\ \mathrm{f}) & = & \mathit{return}\ a\\
\quad \mathit{observe}\ ((\mathit{Return}\ a\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{observe}\ (k\ a\ {:}{\mid}\ \mathrm{f})\\
\quad \mathit{observe}\ (((m\ {:}{\gg\!=}\ k_1)\ {:}{\gg\!=}\ k_2)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{observe}\ ((m\ {:}{\gg\!=}\ (\lambda a \to k_1\ a\ {:}{\gg\!=}\ k_2))\ {:}{\mid}\ \mathrm{f})\\
\quad \mathit{observe}\ ((\mathit{False}\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{observe}\ \mathrm{f}\\
\quad \mathit{observe}\ (((m\ {:}{\mid}\ n)\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{observe}\ ((m\ {:}{\gg\!=}\ k)\ {:}{\mid}\ ((n\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}))\\
\quad \mathit{observe}\ ((\mathit{Promote}\ m\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & m \gg\!= \lambda a \to \mathit{observe}\ (k\ a\ {:}{\mid}\ \mathrm{f})\\
\quad \mathit{observe}\ (\mathit{False}\ {:}{\mid}\ \mathrm{f}) & = & \mathit{observe}\ \mathrm{f}\\
\quad \mathit{observe}\ ((m\ {:}{\mid}\ n)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{observe}\ (m\ {:}{\mid}\ (n\ {:}{\mid}\ \mathrm{f}))\\
\quad \mathit{observe}\ (\mathit{Promote}\ m\ {:}{\mid}\ \mathrm{f}) & = & m\\
\quad \mathit{observe}\ (\mathit{Promote}\ m) & = & m
\end{array}
$$

Figure 6: A term implementation of *BACKTR*.

equations that are concerned with $\omega\ (\bullet \gg\!= c \mid \mathrm{f})$ we see that f appears once in the context $\omega\ \bullet$. Likewise, c is used twice in the context $\omega\ (\bullet\ a\ {:}{\mid}\ \mathrm{f})$. These observations motivate the following specification.

$$
\begin{array}{lll}
\mathit{op}\ \underline{c}\ \underline{\mathrm{f}} = \omega\ (\mathit{op} \gg\!= c \mid \mathrm{f}) & & \\
\quad \Longleftarrow & \underline{\mathrm{f}} = \omega\ \mathrm{f} & (3)\\
\quad \wedge & \forall \mathrm{f}'\ \underline{\mathrm{f}'}.\ (\forall a.\ \underline{c}\ a\ \underline{\mathrm{f}'} = \omega\ (c\ a \mid \mathrm{f}')) \Longleftarrow \underline{\mathrm{f}'} = \omega\ \mathrm{f}' & (4)
\end{array}
$$

The nice thing about Hughes' technique is that mistakes made at this point will be discovered later when the operations are derived. For instance, it may seem unnecessary that $\underline{c}$ is parameterized with $\underline{\mathrm{f}'}$. However, if we simply postulate $\forall a.\ \underline{c}\ a = \omega\ (c\ a \mid \mathrm{f})$, then we will not be able to derive a definition for $(\mid)$. Better still, one can develop the specification above while making the calculations. The derivation of *false*, for instance, motivates assumption (3); the derivation of *return* suggests either $\forall a.\ \underline{c}\ a = \omega\ (c\ a \mid \mathrm{f})$ or assumption (4) and the derivation of $(\mid)$ confirms that (4) is the right choice. The complete derivation appears in Fig. 7. Interestingly, each equation of the specification is invoked exactly once.

It remains to determine the type of the backtracking monad transformer. If we assume that the second parameter, the so-called *failure continuation*, has type $m\ b$, then the first parameter, the so-called *success continuation*, is of type $a \to m\ b \to m\ b$. It follows that the type of the new transformer is $\forall a.(a \to m\ b \to m\ b) \to m\ b \to m\ b$. Again, the answer type is universally quantified. We will see shortly why this is a reasonable choice. Fig. 8 summarizes the implementation.

Reconsider Fig. 7 and note that the derivation of *return*, $(\gg\!=)$, *false*, and $(\mid)$ is completely independent of ω's specification. The laws (O4) and (O5) are only required in the derivation of ω. Only π relies on (O3) which, however, appears to be the only sensible way to observe promoted operations. This suggests that we can define different observations without changing the definitions of the other operations. In other words, we may generalize the specification as follows (here φ is an arbitrary observer function).

$$
\begin{array}{lll}
\mathit{op}\ \underline{c}\ \underline{\mathrm{f}} = \varphi\ (\mathit{op} \gg\!= c \mid \mathrm{f}) & & (5)\\
\quad \Longleftarrow & \underline{\mathrm{f}} = \varphi\ \mathrm{f} & \\
\quad \wedge & \forall \mathrm{f}'\ \underline{\mathrm{f}'}.\ (\forall a.\ \underline{c}\ a\ \underline{\mathrm{f}'} = \varphi\ (c\ a \mid \mathrm{f}')) \Longleftarrow \underline{\mathrm{f}'} = \varphi\ \mathrm{f}' & \\
\quad \wedge & \forall m\ k.\ \varphi\ (\pi\ m \gg\!= k) = m \gg\!= (\varphi \cdot k) &
\end{array}
$$

To illustrate the use of the generalized specification assume that we want to collect all solutions of a nondeterministic computation. To this end we specify an observation *solve* of type $(\mathit{Monad}\ m) \Rightarrow \mathit{BACKTR}\ m\ a \to m\ [a]$:

$$
\begin{array}{rcll}
\mathit{solve}\ \mathit{false} & = & \mathit{return}\ [\,] & \text{(S1)}\\
\mathit{solve}\ (\mathit{return}\ a \mid m) & = & a \lhd \mathit{solve}\ m & \text{(S2)}\\
\mathit{solve}\ (\pi\ m \gg\!= k) & = & m \gg\!= (\mathit{solve} \cdot k), & \text{(S3)}
\end{array}
$$

where $(\lhd)$ is given by

$$
\begin{array}{lcl}
(\lhd) & :: & (\mathit{Monad}\ m) \Rightarrow a \to m\ [a] \to m\ [a]\\
a \lhd \mathit{ms} & = & \mathit{ms} \gg\!= \lambda \mathit{as} \to \mathit{return}\ (a : \mathit{as}).
\end{array}
$$

An implementation for *solve* can be readily derived if we specialize (5) for $c = \mathit{return}$ and $\mathrm{f} = \mathit{false}$. We obtain:

$$
\begin{array}{ll}
\varphi\ \mathit{op} = \mathit{op}\ (\oplus)\ e & \\
\quad \Longleftarrow & \varphi\ \mathit{false} = e\\
\quad \wedge & \forall a\ \mathrm{f}'.\ \varphi\ (\mathit{return}\ a \mid \mathrm{f}') = a \oplus \varphi\ \mathrm{f}'\\
\quad \wedge & \forall m\ k.\ \varphi\ (\pi\ m \gg\!= k) = m \gg\!= (\varphi \cdot k).
\end{array}
$$

$$\begin{array}{ll}
 & (\mathit{return}\ a)\ \underline{c}\ \underline{\mathrm{f}} \\
= & \quad \{ \text{ specification and assumptions (3) \& (4) } \} \\
 & \mathit{observe}\ (\mathit{return}\ a \gg\!= c \mid \mathrm{f}) \\
= & \quad \{ \text{ (M1) } \} \\
 & \mathit{observe}\ (c\ a \mid \mathrm{f}) \\
= & \quad \{ \text{ assumptions (3) \& (4) } \} \\
 & \underline{c}\ a\ \underline{\mathrm{f}}
\end{array}$$

$$\begin{array}{ll}
 & (m \gg\!= k)\ \underline{c}\ \underline{\mathrm{f}} \\
= & \quad \{ \text{ specification and assumptions (3) \& (4) } \} \\
 & \mathit{observe}\ ((m \gg\!= k) \gg\!= c \mid \mathrm{f}) \\
= & \quad \{ \text{ (M3) } \} \\
 & \mathit{observe}\ (m \gg\!= (\lambda a \to k\ a \gg\!= c) \mid \mathrm{f}) \\
= & \quad \{ \text{ specification and assumption (3) } \} \\
 & m\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{observe}\ (k\ a \gg\!= c \mid \mathrm{f}'))\ \underline{\mathrm{f}} \\
= & \quad \{ \text{ specification and assumption (4) } \} \\
 & m\ (\lambda a\ \underline{\mathrm{f}'} \to k\ a\ \underline{c}\ \underline{\mathrm{f}'})\ \underline{\mathrm{f}}
\end{array}$$

$$\begin{array}{ll}
 & \mathit{false}\ \underline{c}\ \underline{\mathrm{f}} \\
= & \quad \{ \text{ specification and assumptions (3) \& (4) } \} \\
 & \mathit{observe}\ (\mathit{false} \gg\!= c \mid \mathrm{f}) \\
= & \quad \{ \text{ (B4) } \} \\
 & \mathit{observe}\ (\mathit{false} \mid \mathrm{f}) \\
= & \quad \{ \text{ (B1) } \} \\
 & \mathit{observe}\ \mathrm{f} \\
= & \quad \{ \text{ assumption (3) } \} \\
 & \underline{\mathrm{f}}
\end{array}$$

$$\begin{array}{ll}
 & (m \mid n)\ \underline{c}\ \underline{\mathrm{f}} \\
= & \quad \{ \text{ specification and assumptions (3) \& (4) } \} \\
 & \mathit{observe}\ ((m \mid n) \gg\!= c \mid \mathrm{f}) \\
= & \quad \{ \text{ (B5) } \} \\
 & \mathit{observe}\ ((m \gg\!= c \mid n \gg\!= c) \mid \mathrm{f}) \\
= & \quad \{ \text{ (B3) } \} \\
 & \mathit{observe}\ (m \gg\!= c \mid (n \gg\!= c \mid \mathrm{f})) \\
= & \quad \{ \text{ specification and assumption (4) } \} \\
 & m\ \underline{c}\ (\mathit{observe}\ (n \gg\!= c \mid \mathrm{f})) \\
= & \quad \{ \text{ specification and assumptions (3) \& (4) } \} \\
 & m\ \underline{c}\ (n\ \underline{c}\ \underline{\mathrm{f}})
\end{array}$$

$$\begin{array}{ll}
 & (\mathit{promote}\ m)\ \underline{c}\ \underline{\mathrm{f}} \\
= & \quad \{ \text{ specification and assumptions (3) \& (4) } \} \\
 & \mathit{observe}\ (\mathit{promote}\ m \gg\!= c \mid \mathrm{f}) \\
= & \quad \{ \text{ (B6) } \} \\
 & \mathit{observe}\ (\mathit{promote}\ m \gg\!= (\lambda a \to c\ a \mid \mathrm{f})) \\
= & \quad \{ \text{ (O3) } \} \\
 & m \gg\!= \lambda a \to \mathit{observe}\ (c\ a \mid \mathrm{f}) \\
= & \quad \{ \text{ assumptions (3) \& (4) } \} \\
 & m \gg\!= \lambda a \to \underline{c}\ a\ \underline{\mathrm{f}}
\end{array}$$

$$\begin{array}{ll}
 & \mathit{observe}\ m \\
= & \quad \{ \text{ (M2) and (B2) } \} \\
 & \mathit{observe}\ (m \gg\!= \mathit{return} \mid \mathit{false}) \\
= & \quad \{ \text{ specification } \} \\
 & m\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{observe}\ (\mathit{return}\ a \mid \mathrm{f}'))\ (\mathit{observe}\ \mathit{false}) \\
= & \quad \{ \text{ (O4) } \} \\
 & m\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{observe}\ (\mathit{return}\ a \mid \mathrm{f}'))\ (\mathit{fail}\ \texttt{"false"}) \\
= & \quad \{ \text{ (O5) } \} \\
 & m\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{return}\ a)\ (\mathit{fail}\ \texttt{"false"})
\end{array}$$

Figure 7: Deriving a context-passing implementation of *BACKTR*.

$$\begin{array}{lll}
\multicolumn{3}{l}{\textbf{type}\ \mathit{BACKTR}\ m\ a} \\
 & = & \forall b.(a \to m\ b \to m\ b) \to m\ b \to m\ b \\
\multicolumn{3}{l}{\textbf{instance}\ (\mathit{Monad}\ m) \Rightarrow \mathit{Monad}\ (\mathit{BACKTR}\ m)\ \textbf{where}} \\
\quad \mathit{return}\ a & = & \lambda \underline{c} \to \underline{c}\ a \\
\quad m \gg\!= k & = & \lambda \underline{c} \to m\ (\lambda a \to k\ a\ \underline{c}) \\
\multicolumn{3}{l}{\textbf{instance}\ (\mathit{Monad}\ m) \Rightarrow \mathit{Backtr}\ (\mathit{BACKTR}\ m)\ \textbf{where}} \\
\quad \mathit{false} & = & \lambda \underline{c} \to \mathit{id} \\
\quad m \mid n & = & \lambda \underline{c} \to m\ \underline{c} \cdot n\ \underline{c} \\
\multicolumn{3}{l}{\textbf{instance}\ \mathit{Transformer}\ \mathit{BACKTR}\ \textbf{where}} \\
\quad \mathit{promote}\ m & = & \lambda \underline{c}\ \underline{\mathrm{f}} \to m \gg\!= \lambda a \to \underline{c}\ a\ \underline{\mathrm{f}} \\
\quad \mathit{observe}\ m & = & m\ (\lambda a\ \underline{\mathrm{f}} \to \mathit{return}\ a)\ (\mathit{fail}\ \texttt{"false"})
\end{array}$$

Figure 8: A context-passing implementation of *BACKTR*.

Consequently, $\mathit{solve}\ \mathit{op} = \mathit{op}\ (\lhd)\ (\mathit{return}\ [\,])$. Now, instead of providing *solve* as an additional observer function we promote it into the backtracking monad.

$$\begin{array}{l}
\mathit{sols} :: (\mathit{Monad}\ m) \Rightarrow \mathit{BACKTR}\ m\ a \to \mathit{BACKTR}\ m\ [a] \\
\mathit{sols}\ m \quad = \quad \pi\ (m\ (\lhd)\ (\mathit{return}\ [\,]))
\end{array}$$

This way we can use the all solution collecting function as if it were a new computational primitive. Since π is a monad morphism, we furthermore know that *sols* satisfies suitable variants of (S1)–(S3). Note that the implementation of *sols* makes non-trivial use of rank-2 types. If we used a variant of *BACKTR* that is parameterized with the answer type, then *sols* cannot be assigned a type $t\ a \to t\ [a]$ for some t.

5. ADDING CONTROL

Let us extend our language by two additional Prolog-like control constructs. The first, called cut and denoted '!', allows us to reduce the search space by dynamically pruning unwanted computation paths. The second, termed *call*, is provided for controlling the effect of cut. Both constructs are introduced as a subclass of *Backtr*.

$$\begin{array}{lll}
\multicolumn{3}{l}{\textbf{class}\ (\mathit{Backtr}\ m) \Rightarrow \mathit{Cut}\ m\ \textbf{where}} \\
\quad ! & :: & m\ () \\
\quad \mathit{cutfalse} & :: & m\ a \\
\quad \mathit{call} & :: & m\ a \to m\ a \\
\quad ! & = & \mathit{return}\ () \mid \mathit{cutfalse} \\
\quad \mathit{cutfalse} & = & ! \gg \mathit{false}
\end{array}$$

The operational reading of '!' and *call* is as follows. The cut succeeds exactly once and returns (). As a side-effect it discards *all* previous alternatives. The operation *call* delimits the effect of cut: *call m* executes *m*; if the cut is invoked in *m*, it discards only the choices made since *m* was called. The class definition contains a third operation, called *cutfalse*, which captures a common programming idiom in Prolog, the so-called cut-fail combination [14].

Note that instances of the class *Cut* must define either '!' or *cutfalse*. The default definitions already employ our knowledge about the properties of the operations, which we shall consider next. We sketch the axiomatization only briefly, for a more in-depth treatment the interested reader is referred to [5]. The cut is characterized by the following three equations.

$$
\begin{aligned}
(\,!\gg m) \shortmid n &= \;!\gg m && \text{(!1)}\\
!\gg (m \shortmid n) &= \;m \shortmid\, !\gg n && \text{(!2)}\\
!\gg \mathit{return}\ () &= \;! && \text{(!3)}
\end{aligned}
$$

The first equation formalizes our intuition that a cut discards *past* choice points, ie alternatives which appear 'above' or to its left. On the other hand, the cut does not affect *future* choice points, ie alternatives which appear to its right. This fact is captured by (!2). Axiom (!3) simply records that cut returns (). An immediate consequence of the axioms is $! = \mathit{return}\ () \shortmid\, ! \gg \mathit{false}$, which explains the default definition of cut. To see why this relation holds replace m by $\mathit{return}\ ()$ and n by false in (!2).

The operation *cutfalse* enjoys algebraic properties which are somewhat easier to remember: *cutfalse* is a left zero of both $(\gg\!=)$ and $(\shortmid)$.

$$
\begin{aligned}
\mathit{cutfalse} \gg\!= k &= \mathit{cutfalse} && \text{(CF1)}\\
\mathit{cutfalse} \shortmid m &= \mathit{cutfalse} && \text{(CF2)}
\end{aligned}
$$

The default definitions use the fact that '!' and *cutfalse* are interdefinable. Likewise, the two sets of axioms are interchangeable. We may either define $\mathit{cutfalse} = \,! \gg \mathit{false}$ and take the equations for '!' as axioms — the laws for *cutfalse* are then simple logical consequences — or vice versa.

Finally, *call* is required to satisfy:

$$
\begin{aligned}
\mathit{call}\ \mathit{false} &= \mathit{false} && \text{(C1)}\\
\mathit{call}\ (\mathit{return}\ a \shortmid m) &= \mathit{return}\ a \shortmid \mathit{call}\ m && \text{(C2)}\\
\mathit{call}\ (\,! \gg m) &= \mathit{call}\ m && \text{(C3)}\\
\mathit{call}\ (m \shortmid \mathit{cutfalse}) &= \mathit{call}\ m && \text{(C4)}\\
\mathit{call}\ (\pi\ m \gg\!= k) &= \pi\ m \gg\!= (\mathit{call} \cdot k). && \text{(C5)}
\end{aligned}
$$

Thus, $\mathit{call}\ m$ behaves essentially like m except that any cut inside m has only local effect. It remains to lay down how the new operations are observed in the underlying monad.

$$
\omega\ (\mathit{call}\ m) = \omega\ m \qquad \text{(O6)}
$$

Note that we need not specify the observation of '!' and *cutfalse* since (C3), (C4), and (O6) imply $\omega\ (\,! \gg m) = \omega\ m$ and $\omega\ (m \shortmid \mathit{cutfalse}) = \omega\ m$.

5.1 A term implementation

The free term implementation faces two problems, one technical and one fundamental. Let us consider the technical problem first. Inspecting the type signature of cut, we find that cut cannot be turned into a constructor, because it does not have the right type. If we define a type, say, $\mathit{CUT}\ m\ a$, then '!' must have exactly this type. Alas, its type signature only allows for a substitution instance, ie $\mathit{CUT}\ m\ ()$. Here, we stumble over the general problem that Haskell's **data** construct is not capable of expressing arbitrary polymorphic term algebras. Fortunately, the axioms save the day. Since '!' can be expressed in terms of *cutfalse* and this operation has a polymorphic type, we turn *cutfalse* into a constructor.

$$
\begin{array}{lcl}
\mathbf{data}\ \mathit{CUT}\ m\ a & = & \mathit{Return}\ a\\
 & \mid & \forall b.\, (\mathit{CUT}\ m\ b) :\!\gg\!= (b \to \mathit{CUT}\ m\ a)\\
 & \mid & \mathit{False}\\
 & \mid & \mathit{CutFalse}\\
 & \mid & \mathit{CUT}\ m\ a :\!\shortmid \mathit{CUT}\ m\ a\\
 & \mid & \mathit{Call}\ (\mathit{CUT}\ m\ a)\\
 & \mid & \mathit{Promote}\ (m\ a)
\end{array}
$$

Turning to the definition of ω we encounter a problem of a more fundamental nature. For a start, we discover that the term $\omega\ (\mathit{call}\ m \gg\!= k)$ cannot be simplified. If we make a further case distinction on m, we end up with $\omega\ (\mathit{call}\ (\mathit{call}\ m \gg\!= k_1) \gg\!= k_2)$ which is not reducible either. The crux is that we have no axiom that specifies the interaction of *call* with $(\gg\!=)$. And rightly so. Each *call* opens a new scope for cut. Hence, we cannot reasonably expect that nested *call*s can be collapsed. This suggests to define two interpreters, one for ω and one for *call*, which means, of course, that the implementation is no longer based on the free term algebra. The resulting code, which is mostly straightforward, appears in Fig. 9. The equations involving *cutfalse* use the fact that *cutfalse* is a left zero of both $(\gg\!=)$ and $(\shortmid)$, and that *call* maps *cutfalse* to *false*. Note that ω falls back on *call* to avoid duplication of code.

5.2 A simplified term implementation

For the sake of completeness, here is the simplified term algebra, which augments the type BACKTR of Sec. 4.2 with an additional constructor for *cutfalse*.

$$
\begin{array}{ll}
\multicolumn{2}{l}{\mathbf{data}\ \mathit{CUT}\ m\ a}\\
\ \ = & \mathit{False}\\
\ \ \mid & \mathit{CutFalse}\\
\ \ \mid & \mathit{Cons}\ a\ (\mathit{CUT}\ m\ a)\\
\ \ \mid & \forall b.\ \mathit{PromoteBind}\ (m\ b)\ (b \to \mathit{CUT}\ m\ a)
\end{array}
$$

In essence, we have lists with two different terminators, *False* and *CutFalse*. Interestingly, exactly this structure (without *PromoteBind*) has been used to give a denotational semantics for Prolog with cut [1], where *cutfalse* and *call* are termed *esc* and *unesc*.

5.3 A context-passing implementation

We have seen that the realization of cut and *call* is more demanding since there is no way to simplify nested invocations of *call*. With regard to the context-passing implementation this means that we must consider an infinite number of possible contexts. Using a grammar-like notation we can characterize the set of all possible contexts as follows.

$$
\mathcal{C} \quad ::= \quad \omega\ (\bullet \gg\!= k \shortmid \mathrm{f}) \mid \mathcal{C}[\mathit{call}\ (\bullet \gg\!= k \shortmid \mathrm{f})]
$$

A context is either simple or of the form $C[\mathit{call}\ (\bullet \gg\!= k \shortmid \mathrm{f})]$ where C is the enclosing context. Thus, contexts are organized in a list- or stack-like fashion. As usual we will represent operations as functions from contexts to observations. The main difference to Sec. 4.3 is that each operation must now consider two different contexts and that the contexts are recursively defined. Note, however, the duality between the term and the context-passing implementation: In Sec. 5.1 we had two interpreters, *call* and ω, and each interpreter had to consider each operation. Here we have two contexts and each operation must consider each context.

Turning to the implementation details we will see that the greatest difficulty is to get the types right. The contexts are represented by a recursive datatype with two constructors: *OBCC* (which is an acronym for <u>o</u>bserve-<u>b</u>ind-<u>c</u>hoice <u>c</u>ontext) and *CBCC* (<u>c</u>all-<u>b</u>ind-<u>c</u>hoice <u>c</u>ontext). The first takes two arguments, the success and the failure continuation, while the second expects three arguments, the two continuations and the representation of the enclosing context. In order to infer their types it is useful to consider the

$$
\begin{array}{lcl}
\mathbf{data}\ \mathit{CUT}\ m\ a & = & \mathit{Return}\ a \\
 & \mid & \forall b.(\mathit{CUT}\ m\ b)\ {:}{\gg\!=}\ (b \to \mathit{CUT}\ m\ a) \\
 & \mid & \mathit{False} \\
 & \mid & \mathit{CutFalse} \\
 & \mid & \mathit{CUT}\ m\ a\ {:}{\mid}\ \mathit{CUT}\ m\ a \\
 & \mid & \mathit{Promote}\ (m\ a) \\
\mathbf{instance}\ \mathit{Cut}\ (\mathit{CUT}\ m)\ \mathbf{where} & & \\
\quad \mathit{cutfalse} & = & \mathit{CutFalse} \\
\quad \mathit{call}\ (\mathit{Return}\ a) & = & \mathit{Return}\ a \\
\quad \mathit{call}\ (\mathit{Return}\ a\ {:}{\gg\!=}\ k) & = & \mathit{call}\ (k\ a) \\
\quad \mathit{call}\ ((m\ {:}{\gg\!=}\ k_1)\ {:}{\gg\!=}\ k_2) & = & \mathit{call}\ (m\ {:}{\gg\!=}\ (\lambda a \to k_1\ a\ {:}{\gg\!=}\ k_2)) \\
\quad \mathit{call}\ (\mathit{False}\ {:}{\gg\!=}\ k) & = & \mathit{False} \\
\quad \mathit{call}\ (\mathit{CutFalse}\ {:}{\gg\!=}\ k) & = & \mathit{False} \\
\quad \mathit{call}\ ((m\ {:}{\mid}\ n)\ {:}{\gg\!=}\ k) & = & \mathit{call}\ ((m\ {:}{\gg\!=}\ k)\ {:}{\mid}\ (n\ {:}{\gg\!=}\ k)) \\
\quad \mathit{call}\ (\mathit{Promote}\ m\ {:}{\gg\!=}\ k) & = & \mathit{Promote}\ m\ {:}{\gg\!=}\ (\mathit{call} \cdot k) \\
\quad \mathit{call}\ \mathit{False} & = & \mathit{False} \\
\quad \mathit{call}\ \mathit{CutFalse} & = & \mathit{False} \\
\quad \mathit{call}\ (\mathit{Return}\ a\ {:}{\mid}\ \mathrm{f}) & = & \mathit{Return}\ a\ {:}{\mid}\ \mathit{call}\ \mathrm{f} \\
\quad \mathit{call}\ ((\mathit{Return}\ a\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{call}\ (k\ a\ {:}{\mid}\ \mathrm{f}) \\
\quad \mathit{call}\ (((m\ {:}{\gg\!=}\ k_1)\ {:}{\gg\!=}\ k_2)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{call}\ ((m\ {:}{\gg\!=}\ (\lambda a \to k_1\ a\ {:}{\gg\!=}\ k_2))\ {:}{\mid}\ \mathrm{f}) \\
\quad \mathit{call}\ ((\mathit{False}\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{call}\ \mathrm{f} \\
\quad \mathit{call}\ ((\mathit{CutFalse}\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{False} \\
\quad \mathit{call}\ (((m\ {:}{\mid}\ n)\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{call}\ ((m\ {:}{\gg\!=}\ k)\ {:}{\mid}\ ((n\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f})) \\
\quad \mathit{call}\ ((\mathit{Promote}\ m\ {:}{\gg\!=}\ k)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{Promote}\ m\ {:}{\gg\!=}\ \lambda a \to \mathit{call}\ (k\ a\ {:}{\mid}\ \mathrm{f}) \\
\quad \mathit{call}\ (\mathit{False}\ {:}{\mid}\ \mathrm{f}) & = & \mathit{call}\ \mathrm{f} \\
\quad \mathit{call}\ (\mathit{CutFalse}\ {:}{\mid}\ \mathrm{f}) & = & \mathit{False} \\
\quad \mathit{call}\ ((m\ {:}{\mid}\ n)\ {:}{\mid}\ \mathrm{f}) & = & \mathit{call}\ (m\ {:}{\mid}\ (n\ {:}{\mid}\ \mathrm{f})) \\
\quad \mathit{call}\ (\mathit{Promote}\ m\ {:}{\mid}\ \mathrm{f}) & = & \mathit{Promote}\ m\ {:}{\mid}\ \mathit{call}\ \mathrm{f} \\
\quad \mathit{call}\ (\mathit{Promote}\ m) & = & \mathit{Promote}\ m \\
\mathbf{instance}\ \mathit{Transformer}\ \mathit{CUT}\ \mathbf{where} & & \\
\quad \mathit{promote} & = & \mathit{Promote} \\
\quad \mathit{observe}\ m & = & \mathit{observe}'\ (\mathit{call}\ m) \\
\mathit{observe}' & :: & (\mathit{Monad}\ m) \Rightarrow \mathit{CUT}\ m\ a \to m\ a \\
\mathit{observe}'\ (\mathit{Return}\ a) & = & \mathit{return}\ a \\
\mathit{observe}'\ (\mathit{Promote}\ m\ {:}{\gg\!=}\ k) & = & m \gg\!= (\mathit{observe}' \cdot k) \\
\mathit{observe}'\ \mathit{False} & = & \mathit{fail}\ \texttt{"false"} \\
\mathit{observe}'\ (\mathit{Return}\ a\ {:}{\mid}\ \mathrm{f}) & = & \mathit{return}\ a \\
\mathit{observe}'\ (\mathit{Promote}\ m\ {:}{\mid}\ \mathrm{f}) & = & m \\
\mathit{observe}'\ (\mathit{Promote}\ m) & = & m
\end{array}
$$

Figure 9: A term implementation of CUT.

specification of the context-passing implementation beforehand. The specification is similar to the one given in Sec. 4.3 except that we have two clauses, one for each context.

$$\begin{aligned}
&\mathit{op}\ (\mathit{OBCC}\ \underline{c}\ \underline{\mathrm{f}}) = \omega\ (\mathit{op} \gg\!= c \shortmid \mathrm{f}) \\
&\quad \Longleftarrow \quad \underline{\mathrm{f}} = \omega\ \mathrm{f} && (4)\\
&\quad \wedge \quad \forall \mathrm{f}'\ \underline{\mathrm{f}'}.\ (\forall a.\ \underline{c}\ a\ \underline{\mathrm{f}'} = \omega\ (c\ a \shortmid \mathrm{f}')) \Longleftarrow \underline{\mathrm{f}'} = \omega\ \mathrm{f}' && (5)\\
&\mathit{op} \cdot \mathit{CBCC}\ \underline{c}\ \underline{\mathrm{f}} = \mathit{call}\ (\mathit{op} \gg\!= c \shortmid \mathrm{f}) \\
&\quad \Longleftarrow \quad \underline{\mathrm{f}} = \mathit{call}\ \mathrm{f} && (6)\\
&\quad \wedge \forall \mathrm{f}'\ \underline{\mathrm{f}'}.\ (\forall a.\ \underline{c}\ a\ \underline{\mathrm{f}'} = \mathit{call}\ (c\ a \shortmid \mathrm{f}')) \Longleftarrow \underline{\mathrm{f}'} = \mathit{call}\ \mathrm{f}' && (7)
\end{aligned}$$

The first clause closely corresponds to the specification of Sec. 4.3. For that reason we may assign the components of $\mathit{OBCC}\ \underline{c}\ \underline{\mathrm{f}}$ the same types: $\underline{\mathrm{f}}$ has type $m\ b$ and $\underline{c}$ has type $a \to m\ b \to m\ b$ where b is the answer type. This implies that the type of contexts must be parameterized with m, a, and b.

$$\mathbf{data}\ \mathcal{C}\ m\ a\ b \quad = \quad \mathit{OBCC}\ (a \to m\ b \to m\ b)\ (m\ b) \mid \ldots$$

The second clause of the specification has essentially the same structure as the first one. The main difference is that the components dwell in the transformed monad rather than in the underlying monad. Furthermore, *CBCC* additionally contains the enclosing context which may have a different type. To illustrate, consider the context $C[\mathit{call}\ (\bullet \gg\!= c \shortmid \mathrm{f})]$ of type $\mathcal{C}\ m\ a\ b$. If we assume that the enclosing context C has type $\mathcal{C}\ m\ i\ b$ — there is no reason to require that C has the same argument type as the entire context, but it must have the same answer type — then f has type $\mathit{CUT}\ m\ i$ and c has type $a \to \mathit{CUT}\ m\ i \to \mathit{CUT}\ m\ i$. This motivates the following definition.

$$\begin{aligned}
\mathbf{data}\ \mathcal{C}\ m\ a\ b \quad &= \quad \mathit{OBCC}\ (a \to m\ b \to m\ b)\ (m\ b)\\
&\mid \quad \forall i. \mathit{CBCC}\ (a \to \mathit{CUT}\ m\ i \to \mathit{CUT}\ m\ i)\\
&\qquad\qquad (\mathit{CUT}\ m\ i)\ (\mathcal{C}\ m\ i\ b)\\
\mathbf{type}\ \mathit{CUT}\ m\ a &= \forall b. \mathcal{C}\ m\ a\ b \to m\ b
\end{aligned}$$

Note that the intermediate type is represented by an existentially quantified variable. The mutually recursive types $\mathcal{C}$ and *CUT* are somewhat mind-boggling as they involve both universal and existential quantification, a combination of features the author has not seen before.

Now that we have the types right, we can address the derivation of the various operations. Except for π the calculations are analogous to those of Sec. 4.3. For $\pi\ m$ we must conduct an inductive proof to show that m propagates through the stack of contexts, ie $(\pi\ m \gg\!= k)\ c = m \gg\!= \lambda a \to k\ a\ c$. The proof is left as an exercise to the reader. To derive cut we reason:

$$\begin{aligned}
&\ !\cdot \mathit{CBCC}\ \underline{c}\ \underline{\mathrm{f}}\\
=&\quad \{\ \text{specification and assumptions (6) \& (7)}\ \}\\
&\ \mathit{call}\ (!\ \gg\!= c \shortmid \mathrm{f})\\
=&\quad \{\ \text{(!3), (M3), and (M1)}\ \}\\
&\ \mathit{call}\ (!\ \gg c\ () \shortmid \mathrm{f})\\
=&\quad \{\ \text{(!1) and (!2)}\ \}\\
&\ \mathit{call}\ (c\ () \shortmid\ !\ \gg \mathit{false})\\
=&\quad \{\ \text{assumption (7)}\ \}\\
&\ \underline{c}\ ()\ (\mathit{call}\ (!\ \gg \mathit{false}))\\
=&\quad \{\ \text{(C3) and (C1)}\ \}\\
&\ \underline{c}\ ()\ \mathit{false}.
\end{aligned}$$

The derivation for the context *OBCC* proceeds in an analogous fashion. For *call* we obtain:

$$\begin{aligned}
&\ \mathit{call}\ m\\
=&\quad \{\ \text{(M2) and (B2)}\ \}\\
&\ \mathit{call}\ (m \gg\!= \mathit{return} \shortmid \mathit{false})\\
=&\quad \{\ \text{specification}\ \}\\
&\ m \cdot \mathit{CBCC}\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{call}\ (\mathit{return}\ a \shortmid \mathrm{f}'))\ (\mathit{call}\ \mathit{false})\\
=&\quad \{\ \text{(C1) and (C2)}\ \}\\
&\ m \cdot \mathit{CBCC}\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{return}\ a \shortmid \mathit{call}\ \mathrm{f}')\ \mathit{false}\\
=&\quad \{\ \underline{\mathrm{f}'} = \mathit{call}\ \mathrm{f}'\ \}\\
&\ m \cdot \mathit{CBCC}\ (\lambda a\ \underline{\mathrm{f}'} \to \mathit{return}\ a \shortmid \underline{\mathrm{f}'})\ \mathit{false}\\
=&\quad \{\ \text{definition } \mathit{cons}\ \}\\
&\ m \cdot \mathit{CBCC}\ \mathit{cons}\ \mathit{false}.
\end{aligned}$$

Thus, *call* installs a new context with *cons* and *fail* as the initial failure continuations. The complete implementation appears in Fig. 10. Note that most of the monad operations *pattern match on the context*. This fact sets the implementation apart from *continuation passing style* (CPS), where the context is an anonymous function that cannot be inspected. By contrast, CPS-based implementations [3, 10] use three continuations (a success, a failure, and a cut continuation).

6. CONCLUSION

Naturally, most of the credit goes to J. Hughes for introducing two wonderful techniques for deriving programs from their specification. Many of the calculations given in this paper already appear in [6], albeit specialized to monads. However, the step from monads to monad transformers is not a big one and this is one of the pleasant findings. To be able to derive an implementation of Prolog's control core from a given axiomatization is quite remarkable. We have furthermore applied the techniques to derive state monad transformers, *STATE*, and exception monad transformers, *EXC*. In both cases the techniques worked well.

Some work remains to be done though. We did not address the problem of promotion in general. It is well known that different combinations of transformers generally lead to different semantics of the operations involved. For instance, composing *STATE* with *BACKTR* yields a backtracking monad with a backtrackable state, which is characterized as follows.

$$\begin{aligned}
\mathit{store}\ s \gg \mathit{false} \quad &= \quad \mathit{false}\\
\mathit{store}\ s \gg (m \shortmid n) \quad &= \quad \mathit{store}\ s \gg m \shortmid \mathit{store}\ s \gg n
\end{aligned}$$

Reversing the order of the two transformers results in a global state, which enjoys a different axiomatization.

$$\mathit{store}\ s \gg (m \shortmid n) \quad = \quad \mathit{store}\ s \gg m \shortmid n$$

For both variants it is straightforward to derive an implementation from the corresponding specification — in the first case $(\shortmid)$ is promoted through *STATE*, in the second case *store* is promoted through *BACKTR*. Unfortunately, some harder cases remain, where the author has not been able to *derive* a promotion in a satisfying way. The problematic operations are, in general, those where the interaction with $(\gg\!=)$ is not explicitly specified. For instance, it is not clear how to derive the promotion of *call* through the state monad transformer.

```
data Ctx m a b  =  OBCC (a → m b → m b) (m b)
                |  ∀i.CBCC (a → CUT m i → CUT m i) (CUT m i) (Ctx m i b)
type CUT m a    =  ∀b.Ctx m a b → m b
instance (Monad m) ⇒ Monad (CUT m) where
    return a    =  λctx₀ → case ctx₀ of OBCC c f → c a f
                                        CBCC c f ctx → c a f ctx
    m ≫= k      =  λctx₀ → case ctx₀ of OBCC c f → m (OBCC (λa f′ → k a (OBCC c f′)) f)
                                        CBCC c f ctx → m (CBCC (λa f′ → k a · CBCC c f′) f ctx)
instance (Monad m) ⇒ Backtr (CUT m) where
    false       =  λctx₀ → case ctx₀ of OBCC c f → f
                                        CBCC c f ctx → f ctx
    m | n       =  λctx₀ → case ctx₀ of OBCC c f → m (OBCC c (n (OBCC c f)))
                                        CBCC c f ctx → m (CBCC c (n · CBCC c f) ctx)
instance (Monad m) ⇒ Cut (CUT m) where
    !           =  λctx₀ → case ctx₀ of OBCC c f → c () (fail "false")
                                        CBCC c f ctx → c () false ctx
    call m      =  λctx₀ → m (CBCC cons false ctx₀)
instance Transformer CUT where
    promote m   =  λctx₀ → case ctx₀ of OBCC c f → m ≫= λa → c a f
                                        CBCC c f ctx → m ≫= λa → c a f ctx
    observe m   =  m (OBCC (λa f → return a) (fail "false"))
```

Figure 10: A context-passing implementation of *CUT*.

7. ACKNOWLEDGMENTS

I would like to thank four anonymous referees for their valuable comments.

8. REFERENCES

[1] M. Billaud. Simple operational and denotational semantics for Prolog with cut. *Theoretical Computer Science*, 71(2):193–208, March 1990.

[2] R. Bird. *Introduction to Functional Programming using Haskell*. Prentice Hall Europe, London, 2nd edition, 1998.

[3] A. de Bruin and E. de Vink. Continuation semantics for prolog with cut. In J. Díaz and F. Orejas, editors, *Proceedings of the International Joint Conference on Theory and Practice of Software Development : Vol. 1*, LNCS 351, pages 178–192. Springer-Verlag, 1989.

[4] N. Dershowitz and J.-P. Jouannaud. Rewrite systems. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, chapter 6, pages 243–320. Elsevier Science Publishers B.V. (North Holland), 1990.

[5] R. Hinze. Prolog's control constructs in a functional setting — Axioms and implementation. *International Journal of Foundations of Computer Science*, 2000. To appear.

[6] J. Hughes. The design of a pretty-printing library. In J. Jeuring and E. Meijer, editors, *Advanced Functional Programming*, LNCS 925, pages 53–96. Springer-Verlag, 1995.

[7] M. P. Jones and L. Duponcheel. Composing monads. Technical Report YALEU/DCS/RR-1004, Department of Computer Science, Yale University, December 1993.

[8] K. Läufer and M. Odersky. An extension of ML with first-class abstract types. In *Proceedings of the 1992 ACM Workshop on ML and its Applications, San Francisco, California*, pages 78–91. ACM-Press, 1992.

[9] S. Liang, P. Hudak, and M. Jones. Monad transformers and modular interpreters. In *Proceedings of the 21st ACM Symposium on Principles of Programming Languages, San Francisco, California*, pages 333–343. ACM-Press, 1995.

[10] E. Meijer. *Calculating Compilers*. PhD thesis, Nijmegen University, 1992.

[11] E. Moggi. An abstract view of programming languages. Technical Report ECS-LFCS-90-113, Department of Computer Science, Edinburgh University, 1990.

[12] E. Moggi. Notions of computation and monads. *Information and Computation*, 93(1):55–92, 1991.

[13] S. Peyton Jones and J. Hughes, editors. *Haskell 98 — A Non-strict, Purely Functional Language*, February 1999. Available from `http://www.haskell.org/definition/`.

[14] L. Sterling and E. Shapiro. *The Art of Prolog: Advanced Programming Techniques*. The MIT Press, 1986.

[15] P. Wadler. Comprehending monads. In *Proceedings of the 1990 ACM Conference on LISP and Functional Programming, Nice*, pages 61–78. ACM-Press, 1990.

[16] P. Wadler. The essence of functional programming. In *Proceedings of the 19th ACM Symposium on Principles of Programming Languages, Sante Fe, New Mexico*, pages 1–14. ACM-Press, 1992.

[17] P. Wadler. Monads for functional programming. In J. Jeuring and E. Meijer, editors, *Advanced Functional Programming*, LNCS 925, pages 24–52. Springer-Verlag, 1995.