Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

Ochre: A Dependently Typed Systems Programming Language

Author:
Charlie Lidbury

Supervisor(s):
Steffen van Bakel
Nicolas Wu

June 11, 2024

Submitted in partial fulfillment of the requirements for the MEng Computing of Imperial College London

Abstract

This research presents Ochre, a dependently typed, low-level systems language. In Ochre, programmers can use the type system to prove stronger properties about their programs than they can in non-dependently typed languages such as Rust or Haskell. Ochre also gives programmers low-level enough control over their programs to be able to express efficient in-place algorithms and control the memory layout of user-defined data structures, which makes it a systems language, akin to Rust, C, or C++.

This paper presents the formal semantics of Ochre via λ_{Ochre}, an abstract interpretation over λ_{Ochre}, a concrete interpretation, a proof that the abstract interpretation and the concrete interpretation are consistent, and an implementation of Ochre in the form of an embedding into the Rust programming language.

Acknowledgments

I would like to thank my supervisor Steffen van Bakel for his type system wisdom, relentless skepticism, and for giving me the freedom to explore such a high-risk project with very little bearing on his research. Steffen even involved his son Isaac van Bakel to help us understand RustBelt and Aeneas, prior work which Ochre takes heavy inspiration from.

I would also like to extend as much gratitude as is physically possible to do via Latex to David Davies, a previous master's student of Steffen who has proven invaluable throughout this project. David has taught me crucial things about dependent types, spent days getting into the nitty gritty of my ideas to make sure I'm on track, and, most importantly, given me the confidence in myself I needed to commit to this project.

Last, but in no means least, I would like to thank my mother Kate Darracott. As well as giving birth to me, which has arguably enabled this project even more than the aforementioned, Mum came up with the brilliant name "Ochre", after being told no more than "the syntax is going to look a little bit like Rust's". Despite not knowing what syntax is, or the significance of dependently typed low-level systems programming languages, she may well have had the most visible contribution to this project of anyone.

Ethical Considerations

Much like Wittgenstein, I believe there is an equivalence between ethics and aesthetics; if you do not, here are a few parallels between the two you might find thought-provoking: We do not choose what we deem ethically permissive, much like we do not choose what we find beautiful. Pursuing one's ethical convictions is not a means to an end, it is an end in and of itself, much like aesthetic experiences.

I and many others including cite cite cite, find aesthetic value in problems & concepts turning out to be reduceable to each other and equivalences being drawn between distant domains. Some particularly high-profile instances of this happening include Euler's formula, the Curry-Howard correspondence and the Church-Turing thesis. To a smaller degree, I also think it happened with Rust's borrow checker, in solving memory management they also solved concurrency, iterator invalidation, and a few other problems that plagued imperative languages.

Despite being sufficiently arrogant and pretentious, I know Ochre isn't as singificant or as beautiful as the previously mentioned identities and isomorphisms. But, in the walled garden of my special interests and obsessions, I have found great aesthetic value in the interplay between ownership semantics and dependent types.

From this aesthetic value, and its equivalence to moral value, I conclude that this research is ethically permissible; I hope Imperial's ethical approval process will too.

Contents

Chapter 1

Introduction

(TAKEN FROM AENEAS FOR NOW)

In 2006, exasperated by yet another crash of his building's elevator's firmware, and exhausted after walking up 21 flights of stairs, Graydon Hoare set out to design a new programming language [Hoare, 2022]. The language, soon to be known as Rust, had two goals. First, to be system-oriented, meaning the programmer would deal with references, pointers, and manually manage memory. Second, to be safe, meaning the compiler's static discipline would rule out memory errors such as use-after-free, or arbitrary memory access. Even though the language evolved a great deal since its inception, these two core premises remain today.

Eighteen years later, Rust enjoys a substantial amount of success and has ranked as the most loved programming language for 7 consecutive years on StackOverflow's developer survey [sta, 2021], until they changed the phrasing of the question in 2023 in which it was the most *admired* language. But as the systems community can attest [**?**Lorch et al., 2020; Ferraiuolo et al., 2017; Bhargavan et al., 2017], memory safety is too weak of a property, no matter how remarkable of an achievement Rust is.

We have attempted to prove further properties

1.1 The Problem

This research hopes to develop a type-checker that is capable of type-checking languages that support both mutation and a kind of type called dependent types. It will do this by removing mutation from the code before type checking, so the type checker only has to reason about immutable code.

Dependent types are covered properly in the background section, but for now, it's enough to know they're a feature that allows you to check even more properties than just type safety at compile time. For instance, instead of just being able to say a variable x is an integer, you can say it's an *even* integer, and reject programs like $x := 5$ at compile time, instead of waiting

for them to go wrong at runtime.

This type-checker will support mutation, which is when a variable's value is changed. For instance, when a variable is declared with a value like $x = 2$, then later given a new value like $x := 5$. The most popular languages all support mutation [cite], it's somewhat the (industry) default. Some languages choose to be *immutable* however, which means they do not support mutation. These include Haskell, and almost all languages with dependent types like Agda, Idris, and Coq.

This type-checker is being built to hopefully be used for a larger, more useful language in the future, called Ochre. Ochre which will have both the speed of *systems languages* like C and Rust and the ability to reason about runtime behaviour at compile time of *theorem provers* like Agda and Coq. Exactly what systems languages and theorem provers are is discussed in Chapter 2.4.

For now, I plan on presenting this type-checker in the form of an implementation; however, there is a good argument for focusing more on the theory behind this type-checker, for instance by presenting a set of typing rules or an abstract algorithm. Whether an implementation-heavy or theory-heavy approach is better is an open, and very important question.

1.1.1 Why Is It Hard?

The problem with having these features together in the same language is that a value that another variable's type depends on can be mutated, which changes the *type* of the other variable. Concretely: if we have a variable $x : T$, and another variable $y : F(x)$ whose type depends on x, we can assign a new value to x which in turn changes the type $F(x)$; now y is ill-typed because its type has changed, but not it's value. The programmer could fix this by reassigning y with a new value of type $F(x)$, if this happens before y is ever used, the compiler should be able to identify this interaction as type-safe.

1.2 The Solution

The technique this research presents goes as follows: convert the source code from the programmer, which will contain mutation, into a functionally equivalent (but maybe inefficient) immutable version, which can be dependently type-checked. Once this immutable version has been type-checked, the original mutable version can be executed, with full efficiency granted to it by mutability.

Because this translation has been shown to be behaviour preserving[**?**] we know properties we prove about the immutable version of the programmer's code also hold for the mutable version which will be executed.

1.3 Motivation

The main contribution of this research will be progress towards making a language that supports both mutability and dependent types, so the motivation behind this research will be the motivation behind these two features, as well as their combination.

This section refers to technical concepts that haven't been explained yet, such as dependent types. The reader is advised to refer to Chapter 2 if they find concepts being referenced that they do not understand.

1.3.1 Mutation

This section argues why one would want mutation in a programming language.

Performance

Some data structures and operations, such as hash maps and their $O(1)$ access/modification, need to modify data in place to be efficiently implemented. Immutable languages like Haskell get around this by performing these mutable operations via unsafe escape hatches and then wrapping those in monads to sequence the immutable operations. However, this often makes mutable code harder to maintain and harder for beginners to understand. For instance, to operate on two hash maps at the same time, you would have to be operating within multiple monads simultaneously, which involves monad transformers or effect types, a much more advanced skillset than what would be required to do the same in Python.

This has widespread effects on the data structures programmers use, and how they structure their programs. Often programmers in immutable languages will simply switch to data structures that don't perform as well but are easier to use in a pure-functional context, like tree-based maps and cons lists instead of hash-maps and vectors.

The performance of explicit mutation can also be easier to reason about. For instance, the Rust code which increments every value in a list of integers doesn't perform any allocations: `for x in xs.iter_mut() { x += 1 }`; whereas the Haskell equivalent looks like it allocates a whole new list, and relies on compiler optimizations to be efficient: `map (+1) xs`. In fact, in this example, Haskell does not do the update in-place and instead allocates a new list in case the old one is being referred to somewhere else. Languages like Koka

Usability

Some algorithms are best thought of in terms of mutable operations, and new programmers especially tend to write stuff mutably. By embracing this in the language design, we can come to the user instead of making the user come to us.

Since the CPU is natively works on mutable operations, if you want control over what the CPU does, which you do if you want to extract all the performance you can from it, you want the language to have graceful support for mutation.

The Immutability Argument

Proponents of immutability argue immutability helps you reason about your program; since there are no side effects of function calls, you cannot be tripped up by side effects you didn't see coming.

I think this correctly identifies that aliased mutation is bad, but goes too far by removing all mutation. In languages like Rust, only one *mutable* reference can exist to any given memory location, which is needed to write to that memory. This gives you most of the benefits of mutation while avoiding the uncontrolled side effects.

Popularity

The majority is often wrong, but it's a good sign if significant proportions of the industry agree on something. In the last quarter of 2023, at least 97.24% of all committed code was written in a language with mutation [cite: GitHut]. At the very least this shows that people like languages with mutability, even if they are wrong to do so.

1.3.2 Dependent Types

This section argues why one would want dependent types in a programming language.

Formal Verification

Dependent types are one of the ways to mechanize logical reasoning, which allows you to reason about the correctness of your programs. For instance, a program that sorts lists should have (amongst other things) the property that it always outputs a list with ascending items. In a language with dependent types, you can make the type of a function express the fact that not only will it return a list of integers, but that it will be a sorted list of integers.

The goal of Ochre, the language this research is done in the name of, is to enable formal verification of low-level systems code. There are other ways to do formal verification, but this is a popular and natural one.

Usability

Dependent types are a notoriously difficult feature to learn and reason about, and their ergonomics are underexplored due to them only being used in very niche, academic languages. However, I think if you're not using them for their extra power, they can be just as ergonomic as typical type systems. In this sense, if the language is designed correctly, you only pay for what you use.

1.3.3 Mutation + Dependent Types

This section explains why mutability and dependent types combine to form more than the sum of their parts.

If you use the mutability to make the language high performance, you can use mutability and dependent types to do formal verification of high performance code. This is a common combination of requirements because they both occur when software is extremely widespread and has very high budget.

1.3.4 This Particular Method

This section explains what advantages this particular method has over other combinations of mutability and dependent types, such as ATS, Magmide, and Low*.

This type checker allows the types and mutable values to be unusually close. In ATS for instance there are basically two separate languages: a dependently typed compile time language and a mutable run-time language. This creates lots of overhead manually linking the two together. For instance, $x : int(y)$ means an integer x with value y. In compile time contexts, you use y to refer to the value, in runtime contexts you use x. I hope to remove the need for this distinction.

Chapter 2

Background

2.1 Dependent Types

2.2 Rust

Rust is a modern programming language that offers a unique combination of strong (memory) safety guarantees and bare-metal performance. Rust innovates in other areas relevant to software engineering, but for this research performance and safety are the two key features which will be built upon.

Performance

Rust is a fast language. Its performance is roughly equivalent to that of C and C++ Rus [b], which are generally accepted as the benchmark of language performance. Rust has enduring performance problems Are [2022], but it is fair to say that on the whole there aren't major performance differences between the fastest languages. The fastest programming languages have more or less hit a ceiling of performance, with no major improvements in speed even since Fortran Gcc which dates back to 1957 [Wilson and Clark, 2001, p. 16].

Making a fast programming language is more about removing slow features than it is about introducing ones that explicitly help performance. Languages like Haskell and Java automatically handle memory allocation and deallocation at the cost of having to have a garbage collector that periodically scans the heap and deallocates inaccessible objects; this is an example of a feature that reduces performance.

Rust is a fast language because it doesn't have a runtime or garbage collector, and has an efficient memory layout. In languages like Haskell or Java, almost all data is heap-allocated and deallocated automatically via a

To generate optimal code, systems languages let the programmer manage their memory, and choose memory layouts. In doing so, they typically sacrifice the memory safety guarantees higher-level languages make due to not being able to check the programmer has managed their memory correctly, this is the case in C and C++. Rust uses a concept called *ownership* to recover these memory safety guarantees while still giving the programmer sufficient control to match C and C++'s performance.

Ownership

Ownership is a set of rules that govern how a Rust program manages memory. All programs have to manage the way they use a computer's memory while running. Some languages have garbage collection that regularly looks for no-longer-used memory as the program runs; in other languages, the programmer must explicitly allocate and free the memory. Rust uses a third approach: memory is managed through a system of ownership with a set of rules that the compiler checks. If any of the rules are violated, the program won't compile. None of the features of ownership will slow down your program while it's running. [1]

There are three rules associated with ownership in Rust:

- Each value in Rust has an owner.
- There can only be one owner at a time.
- When the owner goes out of scope, the value will be dropped.

Borrowing And The Borrow Checker

A consequence of only being able to have one owner of any given value at a time is that passing a value to a function invalidates the variable that used to hold that value. This is referred to as the ownership *moving*. For instance:

```
let x = Box::new(5);
f(x); // Ownership of x passed to f
g(x); // Invalid, we no longer have ownership of x
```

To get around this we could get the functions to give ownership back to us when they return, but this is very syntax-heavy. Rust uses a concept called borrowing in this scenario, which allows you to temporarily give a function access to a value, without giving it ownership. The above example would be done like so:

```
let x = Box::new(5);
f(&x);
g(&x); // Now works
```

[1] Paragraph taken from the Rust Book https://doc.rust-lang.org/book/ch04-01-what-is-ownership.html which I highly recommend for a deeper explanation of ownership.

Here, `&x` denotes a *reference* to `x`. At runtime, this is represented as a pointer. There are two different types of references in Rust: immutable references, denoted by `&T`, and mutable references denoted by `&mut T`. For any given value, you can either hold a single mutable reference or n immutable references, but never both at the same time. This is called the aliasing xor (exclusive or) constraint, or AXM for short.

The borrow checker keeps track of when these references exist to ensure AXM is being upheld. To do this the programmer must annotate references with lifetime annotations, so the compiler has the information of how long the programmer intends each reference to last. Checking these lifetimes overlap in compatible ways is the job of the borrow checker.

2.3 Aeneas & The LLBC

2.4 Prerequisite Concepts

This section explains the concepts required to understand this research.

2.4.1 Mutability

Mutability is when the value of a variable can change at runtime. For instance in Rust, `let mut x = 5; x = 6;` first assigns the value 5 to the variable x, then updates it to 6, which means the value of x depends on the point within the programs execution. This becomes more relevant when you have large objects that get passed around your program, like `let mut v = Vec::new(); v.push(1); v.push(2);` which makes a resizable array on the heap, then pushes 1 and 2 to it.

In Rust to make a variable mutable you must annotate its definition with `mut`, but in most languages, it is just always enabled, like in C `int x = 5; x = 6;` works.

2.4.2 Dependent Types

A dependent type is a type that can change based on the value of another variable in the program. For instance, you might have a variable y which is sometimes an integer, and sometimes a boolean, depending on the value of another variable, x.

When discussing dependent types, there are two important dependent type constructors: Σ and Π. They're usually referenced together because they're roughly equivalent; the dual of Σ types are Π types and visa versa, which apparently means something to category theorists. In the following, I use $Vec(\mathbb{Z}, n)$ to denote the type of an n-tuple of integers, i.e. $(1, 2, 3) : Vec(\mathbb{Z}, 3)$.

- **Dependent Functions** (Π Types) - A dependent function is one whose return type depends on the input value. For instance, you could define a function f which takes a natural n, and returns n copies of 42 in a tuple i.e. $f(3) = (42, 42, 42)$. f's type would be denoted as $f : (\mathbf{n} : \mathbb{N}) \to Vec(\mathbb{Z}, \mathbf{n})$ in Agda/Ochre syntax, or $f : \Pi_{\mathbf{n}:\mathbb{N}} Vec(\mathbb{Z}, \mathbf{n})$ in a more formal mathematical context.
- **Dependent Pairs** (Σ Types) - A dependent pair is a pair where the type of the right element depends on the value of the left element. For instance, you could define a pair p which holds a natural n and a n-tuple of integers i.e. $p = (3, (42, 42, 42))$. p's type would be denoted as $p : (\mathbf{n} : \mathbb{N}, Vec(\mathbb{Z}, \mathbf{n}))$ in Agda/Ochre syntax, or $p : \Sigma_{\mathbf{n}:\mathbb{N}} Vec(\mathbb{Z}, \mathbf{n})$ in a more formal mathematical context.

A language supports dependent types if it can type-check objects like the aforementioned f and s. Just allowing them to exist is not enough. For instance, Python is not dependently typed just because a function's return type can depend on its input, because its type checker doesn't reject programs when you do this wrong. f can be typed in Agda, a dependently typed language with $f : (n : \mathbb{N}) \to Vec(\mathbb{Z}, n)$ but has no valid type in Haskell, which doesn't support dependent types.

2.4.3 Formal Verification with Dependent Types

While dependent types can be nice to have by themselves, a large part of their motivation is using them to perform formal verification.

If you are willing to accept that dependent types can be used to perform formal verification, you do not need to understand how dependent types can be used for logical reasoning: none of this information will be used since the goal of this research is not to perform formal verification, it's just to do dependent type checking.

Readers who are nonetheless interested are invited to read Appendix A.1.

2.4.4 Rust

The mutable $\to$ immutable translation this research relies on requires lifetime annotations to work. While ownership and lifetimes are standalone concepts, their only real-world use case so far has been memory management in the Rust programming language. This section explains these concepts in the context of Rust.

2.4.5 Mutable $\to$ Immutable Translation

To reason about and type-check the mutable code from the programmer, the type checker this research presents translates the source code into an immutable version, as outlined in

Section 1.2.

The crux of this translation is the observation that **a function that mutates a value can be replaced by one that instead returns the new value**. I.e. if the programmer writes a function with type `&mut i32 -> ()`, it can be replaced by `i32 -> i32`. Which would then be used like this:

Listing 2.1: Original

```
let mut x = 5;
f(&mut x); // Mutates x
```

Listing 2.2: Translated

```
let x = 5;
let x = f(x); // Re-defines x
```

The complexity of this translation comes in handling all language constructs in the general case, for instance, if statements need to return the values they edit. Like so:

Listing 2.3: Original

```
let mut x = 5;
if x > 3 {
  x = x + 1; // Mutation
}
```

Listing 2.4: Translated

```
let x = 5;
let x = if x > 3 {
  x + 1
} else {
  x
};
```

This quickly gets complicated when you start to use more advanced features like for loops and functions which return mutable references [2]. So much so safe Rust isn't even entirely covered by the two main attempts at this translation Electrolysis [**?**] and Aeneas [Ho and Protzenko, 2022] [3]. In this research I don't intend to support any constructs not already supported by either of these prior works, so I can use the translation algorithms they have already developed.

2.5 Related Work

Related work comes under two main categories: research which works towards combining mutability with dependent types, and more general work which works towards formal verification of low level code.

[2] See [Ho and Protzenko, 2022] Chapter 2 *Aeneas and its Functional Translation, by Example* for explanation of returning mutable references. (Search for "Returning a Mutable Borrow, and a Backward Function") for the exact paragraph.

[3] See Figure 14 of [Ho and Protzenko, 2022] for a table showing roughly which features are covered by Aeneas/Electrolysis, and see https://kha.github.io/electrolysis/ for exact Rust coverage for Electrolysis.

2.5.1 Languages with Mutability and Dependent Types

ATS

ATS [**?**] is the most mature systems programming language to date, with work dating back to 2002 [ATS, b]. As its website states, it is a *statically typed programming language that unifies implementation with formal specification* [ATS, a].

It's more or less an eagerly evaluated functional language like OCaml, but with functions in the standard library that manipulate pointers, like `ptr_get0` and `ptr_set0` which read and write from the heap respectively. To read or write to a location in memory, you must have a token that represents your ownership of the memory, called a *view*.

For instance, the `ptr_get0` function has the type $\{l : addr\}(T@l|ptr(l)) \rightarrow (T@l|T)$ where

- $\{l : addr\}$ means for all memory addresses, l
- $|$ is the pair type constructor
- $T@l$ means ownership of a value of type T, at location l. Since it is both an input and an output, this function is only *borrowing* ownership.
- $ptr(l)$ means a pointer pointing to location l. Since it can only point at location l, it is a singleton type. This is used to convert the static compile-time variable l into an assertion about the runtime argument.

So overall, this type reads "for all memory addresses l, the function borrows ownership of location l, and turns a pointer to location l into a value of type T".

This necessity to manually pass ownership around introduces a lot of administrative overhead to ATS, which is one of the reasons it is a notoriously hard language to learn/use. ATS introduces syntactic shorthand for these things which you can use in simple cases to clean things up, but still requires this proof passing in many cases which would be dealt with automatically by Rust's borrow checker.

Over the years several versions of ATS have been built, with interesting differences in approach. The current version, ATS2 has only a dependent type-checker, whereas the in-progress ATS3 uses both a conventional ML-like type-checker, as well as a dependent type-checker, and approach that the author of ATS himself developed in separate research, from which ATS3 gets its full name, ATS/Xanadu.

Magmide

The goal of Magmide [**?**] is to "create a programming language capable of making formal verification and provably correct software practical and mainstream". Currently, Magmide is

unimplemented, and there are barely even code snippets of it. However, there is extensive design documentation in which the author Blaine Hansen lays out the compiler architecture he intends to use, which involves two internal representations: *logical* Magmide and *host* Magmide.

- Logical Magmide is a dependently typed lambda calculus of constructions, where to-be-erased types and proofs are constructed.
- Host Magmide is the imperative language that runs on real machines. (Hansen intends on using Rust for this)

I believe this will mean there are two separate languages co-existing on the front end, much like the separation between type-level objects and value-level objects in a language like Haskell.

I suspect this will cause a similar situation to what you see in ATS where for each variable you care about you have two versions, a compile-time one and a runtime one, but it's hard to tell because of the lack of code examples.

Low*

Low*[**?**] is a subset of another language, F*, which can be extracted into C via a transpiler called KreMLin. It has achieved impressive results, mostly at Microsoft Research, where they have used it to implement a formally verified library of modern cryptographic algorithms[**?**] and EverParse

Its set of allowed features is carefully chosen to make this translation possible in the general case, which restricts the ergonomics of the language, it does not support closures, and therefore higher-order programming for example.

It is very much not a pay-for-what-you-use language, to compile anything you must manually manage things like pushing and popping frames on and off the stack, so even if it can achieve impressive results, it's only useful for teams willing to pay the high price which comes with verifying the entire program. This research aims to be better by not requiring any effort from the programmer in the case that they do not wish to use dependent types for their reasoning power.

2.5.2 Embedding Mutability in Languages With Dependent Types

Ynot: Dependent Types for Imperative Programs

Ynot[**?**] is an extension of the Coq proof assistant which allows writing, reasoning about, and extracting higher-order, dependently-typed programs with side-effects including muta-

tion. It does so by defining a monad `ST p A q` which performs an effectful operation, with precondition `p`, postcondition `q` and producing a value of type `A`. They also define another monad, `STSep p A q` which is the same as `ST` except it satisfies the frame rule from separation logic: any part of the heap that isn't referenced by the precondition won't be affected by the computation. This means if you prove properties about a `STSep` computation locally, those proofs still apply even when the computation is put into a different context: this is called compositional reasoning. The Ynot paper presents a formally verified mutable hash table.

Ynot is important foundational work in this area which seems to have inspired many of the other related work here, but is itself not up to the task of verifying low-level code for two reasons:

1. It cannot be used to create performant imperative programs because all mutation occurs through a Coq monad which limits the performance to what you can do in Coq, which is a relatively slow language. This is in contrast to Low*[**?**] for example which is extracted to C, and therefore unrestricted when it comes to performance.

2. To do any verification at all, you must use heap assertions, instead of reasoning about the values directly. This is sometimes needed, like when you're doing aliased mutation (verifying unsafe Rust), but usually not; Aeneas[Ho and Protzenko, 2022] claims to be hugely more productive than its competitors by not requiring heap assertions for safe Rust code.

2.5.3 Formal Verification of Low-Level Code

Low-level code, such as C code can be directly reasoned about by theorem provers like Isabelle, as was done to verify an entire operating system kernel SeL4[**?**]. However, going via C like this has major drawbacks: since the source language is very unsafe, you have a lot of proof obligations. For instance, when reasoning about C you must often prove that a set of pointers do not point to the same location, otherwise mutating the value of one might mutate the others. With Rust references you do not need to do this because the type system prevents you from creating aliased pointers.

Rust Belt

RustBelt[**?**] is a formal model of Rust, including unsafe Rust. Its primary implementation is a Coq framework, Iris[**?**] which allows you to model unsafe Rust code in Coq, and prove it upholds Rust's correctness properties.

I see RustBelt as a great complement to this work in the future: real programs require unsafe code, but you want to avoid having to model your code in a separate proof assistant as little as possible. In Ochre, I imagine the few people who write unsafe code will verify

it with something like RustBelt, while the majority won't have to, but will benefit from the guarantees provided by the verified libraries they use which do.

Chapter 3

Ochre

This chapter introduces the various language constructs of Ochre, at first via intuitive examples, then formally. Each language construct's runtime behavior is discussed, then how it is reasoned about statically, which splits this chapter into 4 sections with the following distinctions:

	Runtime Semantics	Static Analysis
Intuition Building	Section 3.1 Ochre, by Example	Section 3.2 Type & Borrow Checking, by Example
Formal	Section **??** Concrete Interpretation	Section 3.3 Abstract Interpretation

The motivations behind Ochre, alternative design decisions, evaluation, implementation, or reasoning about any properties are all explicit non-goals of this Chapter.

Attribution

With the exception of references, the primitive types in Ochre are from the $\Pi\Sigma$ language presented in Altenkirch et al. [2010], including the representation of algebraic data types.

The mutation & memory management techniques presented are from Rust, including move semantics, references, and the restrictions placed on references.

The novel work presented is the combination of these two features, which requires introducing a new kind of subtyping in which every term is its own type. TypeScript partly does this with literal types, producing results like `5 : 5`, but this research takes this to its logical conclusion where even functions are their own type, and there is almost no distinction between types and terms apart from the requirement that all types are resolved at compile time.

explain double page + hyper links thing

do hyperlinks thing

3.1 Ochre, by Example

This section covers Ochre in a gradual, example-heavy manner, much like programming language tutorials like The Rust Book [Rus, a]. The goal of this section is to build an intuition behind the behavior which the type-checking will later reason about.

Ochre is an impure functional language, composed of expressions that can have side effects.

Basic Language Constructs

The simplest Ochre value is an *atom*. Atoms are constructed with `'`, for example: `'hello` or `'world`[1]. Atoms are an unopinionated primitive type upon which more complex structures can be built.

```
'hello
```

M=N writes the result of evaluating N to M, for example: `x='one` sets `x` to 'one. Declarations are implicit in Ochre (for now); if `x` was in scope previously, `x='one` will bring it into scope, and if it was already in scope, it will mutate it. M;N sequences M, then N. Line comments are opened with `//`.

```
x = 'hello;
x // 'hello
```

References & Mutation

Variables are either modified directly or via a mutable reference. The latter is constructed with `&mut` and eliminated (dereferenced) with `*`.

```
x = 'one;
x = 'two; // mutates x directly
rx = &mut x;
*rx = 'three; // mutates x via a mutable reference
x // 'three
```

Listing 1: Mutation

[1]The runtime representation of an atom is assumed to be the hash of the string after the tick, which makes them constant length. This allows them to be stack-allocated instead of heap-allocated

Whilst a mutable reference to a value exists, that value cannot be read or modified directly, it can only be read or modified via the mutable reference. In Listing 1, the use of `x` on line 5 is not an error despite `rx` existing because is implicitly *dropped* just before the usage of `x`. Because of this implicit drop, `rx` cannot be used after line 5.

In practice, this is intolerably restrictive because it means only one pointer can exist to any value at a time. Like Rust, Ochre solves this by supporting *immutable* references, constructed with `&` and dereferenced with `*`. These allow the programmer to have multiple references to the same value, called *aliasing*. There is a tradeoff that you cannot mutate the referenced value, known as *aliasing xor mutability* (AXM), and it's crucial to how Rust can be converted to pure functional code, or dependently type-checked [Ho and Protzenko, 2022; Ullrich, 2024].

```
1  x = 'one;
2  rx1 = &x;
3  rx2 = &x;
4  x; // 'one
5  *rx1; // 'one
6  *rx2; // 'one
```

Listing 2: The value `'one` can be accessed via `x`, `rx1`, and `rx2` simultaneously

Pairs

M , N constructs the pair of M and N. Pairs are typically surrounded in brackets to make the precedence explicit. M `.0` and M `.1` access the right and left elements of the pair M.

```
1  x = ('one, 'two);
2  x.0; // 'one
3  x.1; // 'two
```

Move Semantics

Ochre uses Rust's ownership semantics to handle manual memory management. Using a value *moves* it, which means it is no longer accessible in the original location. This means you have exclusive access to any value not accessed via an immutable reference. This enables the "whenever a variable goes out of scope, free its associated memory" rule, which is how Rust and Ochre avoid the need for a garbage collector.

Move semantics can lead to some strange results, such as the following program being invalid:

```
1  x = 'one;
```

```
y = x;
x; // error! use of moved value
```

`y = x` moved the value 'one from `x` into `y`, which uninitializes `x`. Moving is granular; you can move components of a pair out of the pair without invalidating the whole pair:

```
x = ('unmoved, 'moved);
y = x.1; // move right component into y
x.0; // 'unmoved
x.1; // error! use of moved value
```

Structural Typing and Type Union

Ochre uses a structural type system. This means a type is entirely defined by the (potentially infinite) set of its inhabitants. This is in contrast to *nominal* typing, where type equivalence depends on the type's name or place of declaration. Take the following type definitions in Rust:

```
struct Foo(i32, i32);
struct Bar(i32, i32);
```

Both `Foo` and `Bar` are types that can be constructed with a pair of integers[2]. In Rust, it would be a type error to pass a `Foo` to a function that expects a `Bar`, because despite holding the same data, they are different types. The equivalent Ochre code would be:

```
Foo = (Int, Int);
Bar = (Int, Int);
```

Unlike in nominally typed languages, an Ochre function which expects a value of type `Foo` as input, can be given a value of type `Bar`. Every identifier you use to refer to a type in Ochre is roughly equivalent to a type *alias* in nominally typed languages like Rust and Haskell.

In Ochre, every value is its own type. So `'one` is of type `'one`, which is expressed in Ochre via colon. Non-singleton types are made up by taking the union of other types, using the `|` operator, like `'a | 'b | 'c`, which can be any of `'a`, `'b`, or `'c`.

```
'a: 'a; // valid
'a: 'a | 'b; // also valid
'c: 'a | 'b; // type error
```

[2] In Rust, `i32` is the type of 32-bit signed integers.

The same goes for references, pairs, and functions (which will be introduced later): the type of a reference is itself a reference, the type of a pair is itself a pair, and the type of a function is itself a function. The only consistent difference between types and terms is types must be statically known, which means they can be erased by runtime.

```
('a, 'b): ('a, 'b); // valid
('a, 'b): ('a | 'b, 'a | 'b); // also valid
```

The $*$ syntax denotes the infinite type/top, the type that contains all values. This is used to represent the concept of no typing information being available. There are three main places where this comes up:

1. Taking the union of two types which don't have a meaningful union, like pairs and atoms. `'a | ('a, 'a): *`.
2. Using it to represent the type of types, which is how you do generic functions. Polymorphic functions are defined by making a function which takes a type as input, and returns a function which uses that type.
3. The type of uninitialised/moved data.

Comptime vs Runtime

Types, just like values, can be assigned to variables for future re-use. However, they must all be statically known, which is enforced by only allowing them to be assigned to *comptime* variables, which start with capital letters. This is similar to how in Haskell types must start with a capital letter, but here the line between types and values is blurred significantly.

```
abPair = ('a | 'b, 'a | 'b); // error! type union can only occur at compile time
ABPair = ('a | 'b, 'a | 'b); // valid
('a, 'b): ABPair;
```

Functions

Functions are defined with an arrow `->` and an optional runtime body surrounded in curly braces. For instance, the identity function over `'true | 'false` is defined as such:

```
Bool = 'true | 'false;
id = (x: Bool) -> Bool { x };
```

If the runtime body is omitted, the function can only be called at compile time, which means it must be written to a comp time variable:

```
Bool = 'true | 'false;
Id = (x: Bool) -> Bool; // valid
id = (x: Bool) -> Bool; // invalid: attempt to assign comptime func to runtime var
```

The only difference between a function body and its return type is that its return type is run at compile time, there is no syntactic difference. For functions you want to run at compile time, syntax after the arrow *is* the function body.

```
Id = x -> x; // Definition of identity which can only be run at comp time
id = x -> x { x }; // Definition of identity which also exists at runtime
```

Case Statements

In Ochre, atoms can be branched on via a case statement. The discriminant of the case statement must be an atom, and there must be exactly one branch for each possible atom. In the future, I plan on adding if and match statements, which will be syntactic sugar for case statements.

```
Bool = 'true | 'false;
not = (b: Bool) -> Bool {
  case b {
    'true => 'false,
    'false => 'true,
  }
};
not('true); // 'false
```

Dependent Pairs

If a pair is being evaluated in a comptime context, the right of a pair can depend on the left. This is done by making the right a function that maps from left to right.

```
Same = (Bool, L -> L); // binds LHS to L, so can be used by right
('true, 'true): Same; // valid
('true, 'false): Same; // error! 'false is not of type 'true

Different = (Bool, L -> case L { 'true => 'false, 'false => 'true});
```

```
    ('true, 'false): Different; // valid
    ('true, 'true): Different; // error!
```

When you union together pairs, it doesn't just union together their left and right and make a new pair, it uses any information it can get from the left pair to more precisely type the right pair.

```
    Same = ('true, 'true) | ('false, 'false);
    // Expanded internally to:
    Same = ('true | 'false, L -> case L { 'true => 'true, 'false => 'false })
```

Listing 3

This makes the union operator precise, taking the union of two types should never produce a type with inhabitants that weren't in either of the types which were unioned together.

If you want to record dependence between the left and right of a pair in a runtime context, you must construct the pair without the dependence, and then use a type constraint to add it back in.

```
    Same = ('true, 'true) | ('false, 'false);
    x = ('true, 'true); // x is a non-dependent pair
    x: Same; // type constraint has made x a dependent pair
```

Type Narrowing

If the right of a pair depends on the left, and then you find something out about the left, you should in turn find something out about the right. This is done in Ochre via type *narrowing*. In the below example, we define a function `f`, and within `f` we know that the left and right of our pair `p` are the same (using the definition in Listing 3). When we match on its left with `p.0`, each branch is type-checked with the additional knowledge that we are in that particular branch. This allows the compiler to correctly identify that when matching on the other side of the pair, you only need to have one branch.

```
    Same = ('true, 'true) | ('false, 'false);
    f = (p: Same) -> Bool {
      case p.0 {
        'true => case p.1 { 'true => 'unit }, // p.1: 'true
        'false => case p.1 { 'false => 'unit }, // p.1: 'false
      }
    }
```

Listing 4: Case statements narrow down the type of their discriminant in each branch

Algebraic Data Types

Take the following definition of Peano naturals in Haskell syntax:

```
data Nat = Zero | Succ Nat
```

In Ochre this is represented by a dependent pair. The left of the pair indicates which variant the ADT is in (either zero or successor), and the right contains the payload of that variant. In the zero case, nothing is stored, so the payload is `'unit`, in the successor case, we store the natural that we are the successor of, so our payload is `Nat`.

```
// "manual" ADT encoding
Nat = (T: 'zero | 'succ, case T { 'zero => 'unit, 'succ => Nat });
// idiomatic encoding using type union
Nat = ('zero, 'unit) | ('succ, Nat);
```

By matching on the left, you can determine which variant the ADT is in, then you can access the payload through the right. For instance, this is how would define addition over Peano naturals:

```
Nat = ('zero, 'unit) | ('succ, Nat);
add = (x: Nat, y: Nat) -> Nat {
  case x.0 {
    'zero => y, // 0 + y = y
    'succ => ('succ, add(x.1, y)), // (1 + x) + y = 1 + (x + y)
  }
}
```

Recursion

The definition of `add` above won't compile because of how it does recursion. When type-checking assignments, Ochre looks at the left first to figure out what type the identifiers have. In the case of `Nat` and `add` above there are no type annotations, so it evaluates the assigned value with no extra type information.

If the programmer puts type annotations on the left of an assignment, the compiler knows at least something about the type, so it can evaluate the expression with that knowledge. This isn't required in the definition of `Nat` because you can put anything in a pair, regardless of its type, so the usage of `Nat` on the right was permissible.

In the `add` case, we need to know that `add` has type (Nat, Nat) -> Nat while evaluating the function body, so we can check that `add(x.1, y)` has type `Nat`. To introduce this, add type annotations to the left of the assignment:

```
add: (Nat, Nat) -> Nat = (x: Nat, y: Nat) -> Nat {
  // ...
}
```

This introduces repetition in the types, which we remove by adding the following syntactic sugar for the above:

```
add(x: Nat, y: Nat): Nat = {
  // ...
}
```

3.2 Type & Borrow Checking, by Example

This section aims to give the reader an intuition behind the abstract interpretation used to type-check Ochre. Specifically, it answers two questions: what is the abstract environment? And how do the various syntactic constructs modify it?

The abstract environment is a mapping from identifiers to types, although it can often look like a mapping from identifiers to values because the type of a value like `'true` is $'true$. It stores the types of both runtime and comptime variables, which are distinguished by comptime variables starting with a capital letter.

Throughout this thesis, syntax will be in monospace font `like this`, and abstract values will be in mathematical text $like\,this$.

Basic Language Features

Type-checking an Ochre program always starts with an empty environment, and every time information is gained, it is added to the abstract environment. Like so. The type of every atom `'a` is the singleton set $\{'a\}$, but it is also every superset of that singleton set like $\{'a, 'b\}$.

```
1  x = 'true; // {x ↦ {'true}}
2  y = 'hello; // {x ↦ {'true}, y ↦ {'hello}}
3  x = 'false; // {x ↦ {'false}, y ↦ {'hello}}
```

Listing 5: A series of assignments, and their corresponding effects on the abstract environment.

In the above example, it would be sound for the abstract environment to map x onto $\{'true, 'false\}$, or even $\{'true, 'unrelated\}$, but that would be losing information. The concept of losing typing information will be made explicit later with environment *rearrangements*, but for now, we'll focus on the environment being as precise as possible.

For brevity, we use $'a$ as syntactic sugar for the singleton set $\{'a\}$. This never causes ambiguity because the abstract environment only ever uses atoms in sets, never by themselves.

```
1  x = 'true; // {x ↦ 'true}
2  y = 'hello; // {x ↦ 'true, y ↦ 'hello}
3  x = 'false; // {x ↦ 'false, y ↦ 'hello}
```

Listing 6: Listing 5 but using syntactic sugar for singlton sets of atoms.

When you move a value, it is mapped to $\bot$ in the abstract environment:

```
x = 'hello; // {x ↦ 'hello}
y = x; // {x ↦ ⊥, y ↦ 'hello}
```

References & Mutation

When you construct a reference, the value is *borrowed*. In the case of mutable borrows, this means the value isn't available in the original location, which is represented in the abstract environment as $\mathsf{loan}^m\, l$ where l is the *loan identifier* for this particular loan. We set it to this instead of $\bot$ so we can find it again in the future when we want to terminate the loan. The reference will map to $\mathsf{borrow}^{\mathsf{m}}\, l\, v$ where v is the type of the value being borrowed.

```
x = 'one; // {x ↦ 'one}
rx = &mut x; // {x ↦ loan^m l, rx ↦ borrow^m l 'one}
*rx = 'two; // {x ↦ loan^m l, rx ↦ borrow^m l 'two}
// rx dropped
x; // {x ↦ 'two, rx ↦ ⊥}
```

Listing 7: A reference to a variable being constructed and used for a mutation. When the reference `rx` is dropped, the updated value from the mutable reference is written back to the original variable `x`.

```
x = 'one; // {x ↦ 'one}
rx = &mut x; // {x ↦ loan^m l, rx ↦ borrow^m l 'one}
*rx = 'two; // {x ↦ loan^m l, rx ↦ borrow^m l 'two}
// rx dropped
x; // {x ↦ 'two, rx ↦ ⊥}
```

Listing 8: A reference to a variable being constructed. When the reference is dropped, the updated value from the mutable reference is written back to the original variable.

Mutable references are similar, except the value is also stored on the loan, reflecting the fact that while an immutable loan exists, the value is still available in its original location. Having loan in an environment like this is also used to prevent mutations to a borrowed value.

```
x = 'one; // {x ↦ 'one}
rx = &x; // {x ↦ loan^s l 'one, rx ↦ borrow^s l 'one}
```

Loans can be nested, which is useful when you want to temporarily give a value you have borrowed to something else.

```
x = 'one; // {x ↦ one}
rx1 = &mut x; // {x ↦ loan^m l, rx ↦ borrow^m l 'one}
rx2 = &mut *rx1; // {x ↦ loan^m l, rx ↦ borrow^m l (loan^m l'), rx2 ↦ borrow^m l 'one}
```

Listing 9: A reborrow

When immutable references are re-borrowed, the syntactic representation of the environment grows exponentially.

```
x = 'one; // {x ↦ one}
rx1 = &x; // {x ↦ loan^s l 'one, rx ↦ borrow^s l 'one}
rx2 = &*rx1; // {x ↦ loan^s l (loan^s l ''one), rx ↦ borrow^s l (loan^s l ''one), rx2 ↦ borrow^s l ''one}
rx3 = &*rx2; // {x ↦ loan^s l (loan^s l' (loan^s l'' 'one)), rx ↦ borrow^s l (loan^s l' (loan^s l'' 'one)),
             // rx2 ↦ borrow^s l' (loan^s l'' 'one), rx3 ↦ borrow^s l'' 'one}
```

Listing 10: An immutable re-borrow

This is not a problem for the implementation because the value stored in the loan and the value stored in the borrow are two pointers to the same underlying memory, it can just make working examples out by hand longer.

(Dependent) Pairs

In the abstract environment pairs store the type of the left side, and how to turn the type of the left side into the right, like so: $(\{'true, 'false\}, \texttt{L} \to \texttt{L})$. This reads "The left of the pair is of type $\{'true, 'false\}$, and the right is whatever the left is". This means in the future if the left is narrowed down to be $'true$, the right will be read as $'true$.

Non-dependent pairs are a special case of dependent pairs where the right happens to evaluate to the same type for any given left. A non-dependent pair of booleans would be constructed with `(Bool, Bool)`, which is syntactic sugar for `(Bool, _ -> Bool)`.

```
Bool = 'true | 'false;                // {Bool ↦ {'true, 'false}}
BoolPair = (Bool, Bool);              // {..., BoolPair ↦ ({'true, 'false}, _ → Bool)}
Same = (Bool, L -> L);                // {..., Same ↦ ({'true, 'false}, L → L}
specificPair = ('true, 'true);        // {..., specificPair ↦ ('true, _ → 'true)}
widenedPair = ('true, 'true): Same; // {..., widenedPair ↦ ({'true, 'false}, L → L}
```

Listing 11: Various pair constructions and their respective entries in the abstract environment

Mutation breaks type dependencies across pairs. Once the left of a pair is mutated, the right must be generalized because the data is lost, meaning the programmer will never be able to recover which specific type the right had in the future.

```
1    Same = ('true,  'true )
2         | ('false, 'false);  // {Same ↦ ({'true, 'false}, L → L)}
3    p = ('true, 'true): Same; // {Same ↦ ..., p ↦ ({'true, 'false}, L → L)}
4    p.0 = 'false;             // {Same ↦ ..., p ↦ ('false, _ → ('true | 'false))}
5    p.1 = 'false;             // {Same ↦ ..., p ↦ ('false, _ → 'false)}
6    p: Same;                  // {Same ↦ ..., p ↦ ({'true, 'false}, L → L)}
```

Listing 12: Demonstration of how mutation interacts with dependent pairs. On line 4 when the left of the pair is mutated, the dependence is broken. When the right is mutated to 'false, the pair's type is narrowed down, but it doesn't regain the dependence until the programmer explicitly widens the type on line 6.

Listing 12 depicts `('true, 'true) | ('false, 'false)` being evaluated to ({'true, 'false}, `L` → `L`), which isn't strictly true. Type union between pairs will make the right depend on the left by producing a case statement for each of the possible left atoms, so `('true, 'true) | ('false, 'false)` would instead evaluate to ({'true, 'false}, `L` → `case L { 'true => 'true, 'false => 'false }`). In code examples it often evaluates to the former, to aid readability.

Type Annotations

Sometimes you want to manually manipulate what type the abstract interpretation reads from a piece of syntax. You do this with type annotations like `M:T`. Evaluating a piece of syntax like `M: T` both asserts that type of M is a subtype of T and makes the expression be of type T instead of M.

```
1    x = 'true; // {x ↦ 'true}
2    y = 'true: 'true | 'false; // {..., y ↦ {'true, 'false}}
```

Listing 13: The type annotation has caused type information to be lost: both `x` and `y` are set to 'true in the above code, but the type annotation on `y` has caused the abstract interpretation to only be able to assign the wider type of {'true, 'false}

Comptime vs Runtime

As you will see in Section 3.3, there are large differences in how the abstract interpretation is performed on runtime and comptime terms; however, for the most part, they map very similarly to the abstract environment. Following from the syntax level distinction, an entry in the abstract environment is marked as runtime or comptime by the variable identifier being capitalized or not.

```
1    x = 'one; // {x ↦ 'one}
2    X = 'one; // {..., X ↦ 'one}
```

One place differences do show is that runtime variables can mutate and be moved, whereas comptime values are immutable and can be freely used like values in typical pure functional languages.

This is so the programmer doesn't have to deal with manual memory management of comptime values, which they wouldn't benefit from anyway because all comptime variables are erased by the time the code is executed.

```
x = 'runtime; // {x ↦ 'runtime}
y = x; // {x ↦ ⊥, y ↦ 'runtime}

X = 'comptime; // {..., X ↦ 'comptime}
Y = X; // {..., X ↦ 'comptime, Y ↦ 'comptime}
```

Listing 14: Unlike the runtime variable `x`, which becomes uninitialized after being moved to `y`, the comptime variable `X` remains accessible while the value is simultaneously used by `Y`, as you would expect from languages move semantics like Haskell

Functions

When a function is called, two things need to be calculated at the call site: whether or not the argument the programmer supplied is a subtype of the required argument; and what the return type is given this argument type. To achieve this we store two pieces of syntax, the input syntax and the return type syntax. We store syntax instead of types so the return type can depend on the input type.

```
Bool = 'true | 'false; // {Bool ↦ {'true, 'false}}
id = (b: Bool) -> Bool {
  b // {Bool ↦ ..., id ↦ ⊤, b ↦ {'true, 'false}}
} // {Bool ↦ {'true, 'false}, id ↦ (b: Bool) → Bool}
```

Listing 15: While type checking the body, argument `b` is in the abstract environment. Abstractly the function is two pieces of *syntax*: `(b: Bool)` and `Bool` instead of their respective types which are both {'true, 'false}

Case Statements

In each of the branches of a case statement, the type of the interrogant is narrowed down to a specific atom. This is useful when the case is branching over the left of a dependent pair, because when the left of the pair gets narrowed down, so does the right[3].

Each branch of the case statement will modify the environment in some possibly different

[3]The right only gets narrowed once it is accessed, not immediately when the left is narrowed

way. These are combined into one environment via an environment-wide union operation, which is the output environment.

```
1   f = (b: 'true | 'false) -> 'unit {
2                         // {b ↦ {'true, 'false}}
3     case b {
4       'true => (        // {b ↦ 'true}
5         x = 'hello;     // {b ↦ 'true, x ↦ 'hello}
6       ),
7       'false => (       // {b ↦ 'false}
8         x = 'world;     // {b ↦ 'false, x ↦ 'world}
9         b = 'true;      // {b ↦ 'true, x ↦ 'world}
10      )
11    };                  // {b ↦ 'true, x ↦ {'hello, 'world}}
12  }
```

Listing 16: Each branch of the case statement is abstractly interpreted with `b` narrowed down to a single atom (lines 4 and 7). Both branches modify `x`, but to different values which make their environments different (lines 5 and 8); these different values are unioned together in the final environment to {'hello, 'world}. The false branch happens to mutate `b` back to 'true, which means by the end of *both* branches, b : 'true, which is reflected in the final environment which maps `b` to 'true instead of {'true, 'false}.

Complex Example Programs

And last but not least, here are a few example programs which use several of the previous features together:

3.3 Formalisms

Type-checking is done via an abstract interpretation which takes a piece of Ochre syntax and outputs a type for that syntax while modifying the abstract environment. All interpretation rules take the form $\Omega \vdash M \diamond v \dashv \Omega'$ where $\diamond$ is one of $\{ \dot{\rightsquigarrow}, \overset{(\cdot)}{\rightarrow}, \overset{(\cdot)}{\Rightarrow}, \overset{(\cdot)}{\Leftarrow}, \overset{(\cdot)}{\leftarrow}, \dot{\leftsquigarrow} \}$, M is Ochre syntax, Ω and Ω' are the abstract environments before and after, and v is the type of the value which has been read or written.

Subsection 3.3.1 introduces key concepts that are required to understand the definitions; subsequent subsections define the abstract interpretation. The reader is encouraged to skip around definitions a lot, they are laid out in tables to make finding a particular definition easier, as one might in a repository of code.

3.3.1 Modalities

It is best to conceptualize the arrows as a single interpretation with many modalities. The following arrows are used to denote the different modalities:

	Read	Write
Runtime destructive	$\overset{(\cdot)}{\Rightarrow}$ *move*	$\overset{(\cdot)}{\Leftarrow}$ *write*
Runtime non-destructive	$\overset{(\cdot)}{\rightarrow}$ *read*	$\overset{(\cdot)}{\leftarrow}$ *type narrow*
Comptime	$\dot{\rightsquigarrow}$ *erased read*	$\dot{\leftsquigarrow}$ *erased write*

To summarize the modalities: a dot above the arrow denotes *abstract* interpretation which means it is not executing the code, only type-checking it; squiggly arrows denote the interpretation of terms that are erased at compile time; arrows which point rightwards (away from the term) denote reads, arrows which point leftwards (towards the term) denote writes. These modalities are elaborated on below.

Comptime vs Runtime Modality

There are two distinct types of terms in Ochre: runtime terms, and comptime terms.

Runtime terms are constrained such that they can be executed efficiently on hardware. This involves manual memory management with move semantics, so no garbage collector is required at runtime, and allows the in-place mutation of data structures.

Comptime terms are erased at compile time and are only used to compute types. Because comptime terms only exist during compilation, and not during runtime, inefficient but automatic memory management strategies can be used, such as reference counting. This removes

the need for move semantics, which allows types to be used multiple times without explicit copying.

Because comptime terms do not have move semantics, we cannot have mutation[4], and we do not need immutable references. Not supporting mutation within comptime terms has the added benefit of the programmer not having to reason about the side effects of evaluating types, which happens implicitly in situations including type checking a function call site (see $\langle$*def.* $\overset{(\cdot)}{\Rightarrow}$ *for* $MN\rangle$).

There is no syntactic distinction between comptime and runtime terms because they are so similar, although there could have been because one can always determine whether a term is comptime or runtime given it's position.

Abstract vs Concrete Modality

Comptime interpretations ($\dot{\rightsquigarrow}$, $\dot{\leftsquigarrow}$) are only ever abstract, whereas runtime interpretations ($\overset{(\cdot)}{\Rightarrow}$, $\overset{(\cdot)}{\Leftarrow}$, $\overset{(\cdot)}{\leftarrow}$, $\overset{(\cdot)}{\rightarrow}$) can have an optional dot above them; this dot means "abstract interpret".

Execution and type checking are very closely related in Ochre because execution is just the totally precise version of type checking. They differ only in how they treat type annotations: abstract interpretation will evaluate $M : T$ to the result of evaluating T, and concrete interpretation will evaluate the same syntax to the result of evaluating M.

This guarantees concrete interpretation will output a precise result (singleton type) because the only way to form non-singleton types is via type union, which can only occur on the right hand side of a type annotation, because it's only permitted in comptime terms.

Read vs Write Modality

In typical languages like the λ-calculus, terms are evaluated to a value, which is equivilant to the notion of the read modality in Ochre. The write modality allows you to write a value to a term, which is how variables are brought into scope. You might find it useful to compare reading a variable x with writing to a variable x:

$$\frac{\langle \textit{def.} \overset{(\cdot)}{\Rightarrow} \textit{for}\, x\rangle \qquad \Omega' = \Omega\left[\frac{x \mapsto \top}{x \mapsto v}\right]}{\Omega \vdash x \overset{(\cdot)}{\Rightarrow} m \dashv \Omega'} \qquad \frac{\langle \textit{def.} \overset{(\cdot)}{\Leftarrow} \textit{for}\, x\rangle \qquad \Omega' = \Omega\left[\frac{x \mapsto v}{x \mapsto \top}\right]}{\Omega \vdash x \overset{(\cdot)}{\Leftarrow} v \dashv \Omega'}$$

Figure 3.1: Reading removes a value from the environment, whereas writing adds a value.

[4]The method of combining mutability with dependent types this research uses relies on move semantics, therefore we cannot have mutability on non-move semantics code.

Defining the write operation for more complex pieces of syntax is how several language features are defined, including but not limited to: pattern matching, destructuring, and specifying function arguments.

3.3.2 Syntax

Ochre grammar:

$$
\begin{array}{lll}
S & ::= & \\
 & \quad M & /\!/ \textit{ expression} \\
 & \quad M = N; S & /\!/ \textit{ assignment} \\
 & \quad \texttt{match}\, M\, \{\, \overrightarrow{M' \texttt{ => } S}\, \} & /\!/ \textit{ match statement} \\
 & & \\
T, U, M, N & ::= & /\!/ \textit{ expression} \\
 & \quad x \mid y \mid z & /\!/ \textit{ runtime variable identifier} \\
 & \quad X \mid Y \mid Z & /\!/ \textit{ comptime variable identifier} \\
 & \quad \texttt{'}a & /\!/ \textit{ atom construction} \\
 & \quad M\texttt{,}N & /\!/ \textit{ pair construction} \\
 & \quad M\texttt{.0} & /\!/ \textit{ pair left access} \\
 & \quad M\texttt{.1} & /\!/ \textit{ pair right access} \\
 & \quad \texttt{*}M & /\!/ \textit{ dereference} \\
 & \quad \texttt{\&}M \mid \texttt{\&mut}\, M & /\!/ \textit{ borrow constructor} \\
 & \quad MN & /\!/ \textit{ function application} \\
 & \quad M\, \texttt{->}\, T\, (\texttt{\{}\, N\, \texttt{\}}) & /\!/ \textit{ function definition (optional runtime body)} \\
 & \quad \_ & /\!/ \textit{ uninitialised} \\
 & \quad T \texttt{ | } U & /\!/ \textit{ type union} \\
 & \quad M\texttt{:}\, T & /\!/ \textit{ type constraint} \\
 & \quad v & /\!/ \textit{ type/value}
\end{array}
$$

Figure 3.2: Ochre syntax

Assignments never occur in a terminal position. This avoids the question of what is the return value of an assignment.

Match statements always occur in a terminal position. RustBelt's λ_{Rust} Jung et al. [2018] has the same restriction, but Aeneas' LLBC Ho and Protzenko [2022] does not. If it was permitted for operations to occur after a match statement, the environment *after* the match statement would have to be calculated, which would be the type union of the environments produced by the branches. For this type union operation to be precise over environments, like it is on pairs, we would have to support dependencies between variables. We do not support such dependencies for the sake of simplicity, and thus cannot precisely define environment union.

This restriction does not limit what programs can be represented because a program with

match statements in non-terminal positions can be re-written to one that only has matches in terminal positions. This could be done by moving everything after the match statement into each of the match statement branches:

$$\forall \diamond \{ \overset{(\cdot)}{\Rightarrow}, \dot{\rightsquigarrow} \}. \left[\frac{\Omega \vdash \texttt{match}\ M\ \texttt{\{}\ \overrightarrow{M'\ \texttt{=>}\ S; S'}\ \texttt{\}}\ \diamond t}{\Omega \vdash \texttt{match}\ M\ \texttt{\{}\ \overrightarrow{M'\ \texttt{=>}\ S}\ \texttt{\}}\ ; S' \diamond t} \right]$$

Figure 3.3: A re-write rule which could enable matches in non-terminal positions.

This rewrite rule has not been included in Ochre because I intend to support dependence between variables in the future, and therefore precise environment union, which removes the need for the rewrite. This rewrite rule has the disadvantage of causing an exponential blowup in code size/interpretation derivation size.

Assignment and match statements are the only constructs that can narrow types in the environment, so by having them both in statement (S) instead of expression (M), we guarantee that expression evaluation only ever widens types. This simplifies type-checking expressions.

Types/values can be treated as syntax. This syntactic construction cannot be constructed by the programmer. It is constructed internally within the interpretations to make syntax that always returns the same type/value, such as in $\langle$*def.* $\overset{(\cdot)}{\Rightarrow}$ *for* M . $0\rangle$ where it is used to break the dependence of the pair so the left value can be moved out.

3.3.3 Abstract Environment/Values

The abstract environment, and abstract values:

$$\begin{array}{lll}
\Omega & ::= & \text{// abstract environment (stack)} \\
 & \quad \emptyset & \text{// empty stack} \\
 & \quad \Omega, x \mapsto v & \text{// runtime variable} \\
 & \quad \Omega, X \mapsto v & \text{// comptime variable} \\
 & \quad \Omega, l \mapsto v & \text{// loan restriction} \\
 & & \\
m, n, v, w, t, u & ::= & \text{// type/value} \\
 & \quad \{\overrightarrow{\text{'a}}\} & \text{// atom} \\
 & \quad (v, T \to U) & \text{// pair} \\
 & \quad (T \to U) & \text{// function} \\
 & \quad \mathsf{borrow}^{\mathsf{s}}\, l\, v \mid \mathsf{borrow}^{\mathsf{m}}\, l\, v & \text{// reference} \\
 & \quad \mathsf{loan}^{s}\, l\, v \mid \mathsf{loan}^{m}\, l & \text{// referenced value} \\
 & \quad \top & \text{// top}
\end{array}$$

Figure 3.4: Abstract/Concrete Environment and Types

The Top Type

The $\top$ type is used to denote a lack of typing information. Every type/value is of type $\top$. When you move a value, the previous location is set to $\top$, to denote uninitialized data. When a value has never been writen to before, it's value is $\top$, again, to denote uninitialised data. When a type t depends on another type u, but u has not been narrowed down enough to deduce the type of t, t evaluates to $\top$ to denote the lack of typing information.

Concrete Values

If a type is a singleton type (a type with one inhabitant), it is referred to as a value. For example {'a} is a concrete value/type, but {'a, 'b} is not. Figure 3.6 shows the formal definition of concrete values. Concrete values and non-singleton types share a grammar because rules are typically generic over both and preserve a values concreteness, so combining them avoids the syntatic overhead of introducing an additional modality.

Drop Operation

When an operation is no longer used, it must be dropped. At runtime, dropping a value will free it's associated memory, allowing it to be used for other operations, which is why Ochre doesn't need a garbage collector. Dropping a reference to a value removes the restrictions created by that reference. Drop is defined in Figure 3.6.

Environment Rearrangement

At any point during a program interpretation, whether it be abstract or concrete interpretation, the environment can be *rearranged*, a technique introduced by Aeneas Ho and Protzenko [2022]. Environment rearranges can be inserted before or after any of the interpretation judgements (as defined in Figure 3.5) and can perform one of the following operations:

1. **Allocation** - Before a variable is used, including before it is first written to, it must be mapped to $\top$ in the environment. Allocation takes a variable previously not in the environment, and maps it to $\top$.

2. **Deallocation** - Occasionally typing judgements will assert that a series of operations leave the environment back in its original state (see ⟨*def.* $\overset{(\cdot)}{\Rightarrow}$ *for* M -> T { N }⟩). In order to achieve this variables allocated in that series of operations must be deallocated.

3. **Type Widening** - At any point during the interpretation it is valid to forget typing information. For example: if a value is known to be one of {'a, 'b}, it is valid to now consider it to be one of {'a, 'b, 'c}.

$$\frac{\Omega' = \Omega, xX \mapsto \top}{\Omega \hookrightarrow \Omega'}\ \textsc{Allocation} \qquad \frac{\Omega', xX \mapsto \top = \Omega}{\Omega \hookrightarrow \Omega'}\ \textsc{Deallocation} \qquad \frac{\Omega' = \Omega\left[\frac{x \mapsto v'}{x \mapsto v}\right] \quad \Omega \vdash v \sqsubseteq v'}{\Omega \hookrightarrow \Omega'}\ \textsc{Type-Widen} \qquad \frac{\Omega' = \Omega\left[\frac{x \mapsto \top}{x \mapsto v}\right] \quad \Omega' \vdash \mathsf{drop}\, v \dashv \Omega''}{\Omega \hookrightarrow \Omega''}\ \textsc{Drop}$$

$$\forall \diamond \in \{ \dot{\rightsquigarrow}, \overset{(\cdot)}{\rightarrow}, \overset{(\cdot)}{\Rightarrow}, \overset{(\cdot)}{\Leftarrow}, \overset{(\cdot)}{\leftarrow}, \dot{\rightsquigarrow} \} \left[\frac{\Omega \hookrightarrow \Omega' \quad \Omega' \vdash M \diamond v \dashv \Omega''}{\Omega \vdash M \diamond v \dashv \Omega''}\right] \textsc{Rearrange-Before}$$

$$\forall \diamond \in \{ \dot{\rightsquigarrow}, \overset{(\cdot)}{\rightarrow}, \overset{(\cdot)}{\Rightarrow}, \overset{(\cdot)}{\Leftarrow}, \overset{(\cdot)}{\leftarrow}, \dot{\rightsquigarrow} \} \left[\frac{\Omega \vdash M \diamond v \dashv \Omega' \quad \Omega' \hookrightarrow \Omega''}{\Omega \vdash M \diamond v \dashv \Omega''}\right] \textsc{Rearrange-After}$$

Figure 3.5: Definition of environment rearrangement $\Omega \hookrightarrow \Omega'$. `xX` denotes "either runtime or comptime variable".

4. **Dropping** - Before deallocation, values must be dropped. This is achieved in derivations by inserting rearrangements which drop values.

Environment rearrangement is defined in Figure 3.5.

v	$\Omega \vdash \mathsf{drop}\, v \dashv \Omega'$	$\mathsf{concrete}\, v$
$\{\overrightarrow{\text{'a}}\}$	$\dfrac{}{\Omega \vdash \mathsf{drop}\, \{\overrightarrow{\text{'a}}\}}$	$\dfrac{}{\mathsf{concrete}\, \{\text{'a}\}}$
$(v, T \to U)$	$\dfrac{\Omega \vdash T \dot{\leftsquigarrow} v \dashv \Omega' \quad \Omega' \vdash U \dot{\rightsquigarrow} w \quad \Omega \vdash \mathsf{drop}\, v \dashv \Omega'' \quad \Omega'' \vdash \mathsf{drop}\, w \dashv \Omega'''}{\Omega \vdash \mathsf{drop}\, (v, T \to U) \dashv \Omega'}$	$\dfrac{\mathsf{concrete}\, v \quad \mathsf{concrete}\, (T \to U)}{\mathsf{concrete}\, (v, T \to U)}$
$(T \to U)$	$\dfrac{}{\Omega \vdash \mathsf{drop}\, (T \to U)}$	$\dfrac{\mathsf{runtime}\, T \quad \mathsf{runtime}\, U}{\mathsf{concrete}\, (T \to U)}$
$\mathsf{borrow}^{\mathsf{s}}\, l\, v$	$\dfrac{\Omega' = \Omega \left[\dfrac{v}{\mathsf{loan}^{s}\, l\, v}\right]}{\Omega \vdash \mathsf{drop}\, (\mathsf{borrow}^{\mathsf{s}}\, l\, v) \dashv \Omega'}$	$\dfrac{\mathsf{concrete}\, v}{\mathsf{concrete}\, (\mathsf{borrow}^{\mathsf{s}}\, l\, v)}$
$\mathsf{borrow}^{\mathsf{m}}\, l\, v$	$\dfrac{\Omega' = \Omega \left[\dfrac{v}{\mathsf{loan}^{m}\, l}\right]}{\Omega \vdash \mathsf{drop}\, (\mathsf{borrow}^{\mathsf{m}}\, l\, v) \dashv \Omega'}$	$\dfrac{\mathsf{concrete}\, v}{\mathsf{concrete}\, (\mathsf{borrow}^{\mathsf{m}}\, l\, v)}$
$\mathsf{loan}^{s}\, l\, v$	$\dfrac{\Omega' = \Omega \left[\dfrac{\top}{\mathsf{borrow}^{\mathsf{s}}\, l\, v}\right] \quad \Omega' \vdash \mathsf{drop}\, v \dashv \Omega''}{\Omega \vdash \mathsf{drop}\, (\mathsf{borrow}^{\mathsf{s}}\, l\, v) \dashv \Omega''}$	$\dfrac{\mathsf{concrete}\, v}{\mathsf{concrete}\, (\mathsf{loan}^{s}\, l\, v)}$
$\mathsf{loan}^{m}\, l$	$\dfrac{\Omega' = \Omega \left[\dfrac{\top}{\mathsf{borrow}^{\mathsf{m}}\, l\, v}\right] \quad \Omega' \vdash \mathsf{drop}\, v \dashv \Omega''}{\Omega \vdash \mathsf{drop}\, (\mathsf{borrow}^{\mathsf{m}}\, l\, v) \dashv \Omega''}$	$\dfrac{}{\mathsf{concrete}\, (\mathsf{loan}^{m}\, l)}$

Figure 3.6: Definition of drop and concrete

3.3.4 Typing Judgements

The tables are to aid the reader to locate rules, they do not encode any information themselves.

These definitions are referenced with $\langle$*def.* M *for* $\rightarrow\rangle$, where $\rightarrow$ is the arrow being defined, and M, is the piece of syntax it is being defined for. For example $\langle$*def.* MN *for* $\overset{(\cdot)}{\Rightarrow}\rangle$ refers to the rule for destructively reading a function application.

$$\forall \diamond \in \{ \dot{\rightsquigarrow}, \rightarrow, \Rightarrow \}. \left[\frac{\begin{array}{ll} \Omega \vdash M \diamond m \dashv \Omega' & /\!/ \textit{ hello world} \\ \Omega' \vdash T \dot{\rightsquigarrow} t & /\!/ \textit{ hello world} \\ \Omega' \vdash m \sqsubseteq t & /\!/ \textit{ hello again} \end{array}}{\Omega \vdash M : T \diamond t \dashv \Omega'} \right]$$

	$\overset{\cdot}{\rightsquigarrow}$	$\overset{(\cdot)}{\rightarrow}$	$\overset{(\cdot)}{\Rightarrow}$	$\overset{(\cdot)}{\Leftarrow}$	$\overset{(\cdot)}{\leftarrow}$	$\overset{\cdot}{\rightsquigarrow}$
`'a`	$\forall \diamond \in \{ \overset{\cdot}{\rightsquigarrow}, \overset{(\cdot)}{\rightarrow}, \overset{(\cdot)}{\Rightarrow}, \overset{(\cdot)}{\Leftarrow}, \overset{(\cdot)}{\leftarrow}, \overset{\cdot}{\rightsquigarrow} \}. \left[\dfrac{}{\Omega \vdash \texttt{'a} \diamond {}'a} \right]$					
`_`	$\forall \diamond \in \{ \overset{(\cdot)}{\rightarrow}, \overset{(\cdot)}{\Rightarrow}, \overset{(\cdot)}{\Leftarrow}, \overset{(\cdot)}{\leftarrow} \}. \left[\dfrac{}{\Omega \vdash \_ \diamond \top} \right]$					
`*`	$\dfrac{}{\Omega \vdash * \overset{\cdot}{\rightsquigarrow} \top}$					$\dfrac{}{\Omega \vdash * \overset{\cdot}{\rightsquigarrow} \top}$
`M:T`	$\forall \diamond \in \{ \overset{\cdot}{\rightsquigarrow}, \overset{\cdot}{\rightarrow}, \overset{\cdot}{\Rightarrow} \}. \left[\dfrac{\begin{array}{c}\Omega \vdash M \diamond m \dashv \Omega' \\ \Omega' \vdash T \overset{\cdot}{\rightsquigarrow} t \\ \Omega' \vdash m \sqsubseteq t\end{array}}{\Omega \vdash M : T \diamond t \dashv \Omega'} \right]$				$\forall \diamond \in \{ \overset{\cdot}{\rightsquigarrow}, \overset{\cdot}{\leftarrow}, \overset{\cdot}{\Leftarrow} \}. \left[\dfrac{\begin{array}{c}\Omega \vdash M \diamond m \dashv \Omega' \\ \Omega' \vdash T \overset{\cdot}{\rightsquigarrow} t \\ \Omega' \vdash m \sqsubseteq t\end{array}}{\Omega \vdash M : T \diamond t \dashv \Omega'} \right]$	
`M:T`	$\forall \diamond \in \{ \overset{\cdot}{\rightsquigarrow}, \overset{\cdot}{\rightarrow}, \overset{\cdot}{\Rightarrow} \}. \left[\dfrac{\begin{array}{l}\Omega \vdash M \diamond m \dashv \Omega' \\ \Omega' \vdash T \overset{\cdot}{\rightsquigarrow} t \\ \Omega' \vdash m \sqsubseteq t\end{array}}{\Omega \vdash M : T \diamond t \dashv \Omega'} \right]$				$\forall \diamond \in \{ \overset{\cdot}{\rightsquigarrow}, \overset{\cdot}{\leftarrow}, \overset{\cdot}{\Leftarrow} \}. \left[\dfrac{\begin{array}{c}\Omega \vdash M \diamond m \dashv \Omega' \\ \Omega' \vdash T \overset{\cdot}{\rightsquigarrow} t \\ \Omega' \vdash m \sqsubseteq t\end{array}}{\Omega \vdash M : T \diamond t \dashv \Omega'} \right]$	
	$\forall \diamond \in \{ \rightarrow, \Rightarrow, \Leftarrow, \leftarrow \}. \left[\dfrac{\Omega \vdash M \diamond m \dashv \Omega' \quad /\!/ \textit{ ignore types}}{\Omega \vdash M : T \diamond m \dashv \Omega'} \right]$					

Figure 3.7: everythingtable

M	$\Omega \vdash M \overset{(\cdot)}{\Rrightarrow} v \dashv \Omega'$	$\Omega \vdash M \overset{(\cdot)}{\to} v$	$\Omega \vdash M \dot{\rightsquigarrow} v$
`x` or `X`	$\dfrac{\Omega' = \Omega\left[\dfrac{x \mapsto \top}{x \mapsto v}\right]}{\Omega \vdash x \overset{(\cdot)}{\Rrightarrow} m \dashv \Omega'}$	$\dfrac{x \mapsto v \in \Omega}{\Omega \vdash \texttt{x} \overset{(\cdot)}{\to} v}$	$\dfrac{X \mapsto t \in \Omega}{\Omega \vdash \texttt{X} \dot{\rightsquigarrow} t}$
M, N	$\forall \diamond \in \{\overset{(\cdot)}{\Rrightarrow}, \overset{(\cdot)}{\to}\}. \left[\dfrac{\Omega \vdash M \diamond m \dashv \Omega' \quad \Omega' \vdash N \diamond n \dashv \Omega''}{\Omega \vdash (M,\ N) \diamond (m, \_ \to n) \dashv \Omega''}\right]$		$\dfrac{\Omega \vdash (T,\ \_\ \texttt{->}\ U) \dot{\rightsquigarrow} v}{\Omega \vdash (T,\ U) \dot{\rightsquigarrow} v}$ $\dfrac{\Omega \vdash T \dot{\rightsquigarrow} t \quad \Omega \vdash T' \dot{\leftsquigarrow} t \dashv \Omega' \quad \Omega' \vdash U \dot{\rightsquigarrow} u}{\Omega \vdash (T,\ T'\ \texttt{->}\ U) \dot{\rightsquigarrow} (t, T' \to U)}$
$M.0$	$\dfrac{\Omega \vdash M \overset{(\cdot)}{\Rrightarrow} (v, T \to S_t) \dashv \Omega' \quad \Omega' \vdash T \dot{\leftsquigarrow} v \dashv \Omega'' \quad \Omega'' \vdash S_t \dot{\rightsquigarrow} w \quad \Omega' \vdash M \overset{(\cdot)}{\Leftarrow} (\top, \_ \to w) \dashv \Omega''}{\Omega \vdash M.0 \overset{(\cdot)}{\Rrightarrow} m \dashv \Omega''}$		$\forall \diamond \in \{\overset{(\cdot)}{\to}, \dot{\rightsquigarrow}\}. \left[\dfrac{\Omega \vdash M \diamond (m, \_ \to \_) \dashv \Omega'}{\Omega \vdash M.0 \diamond m \dashv \Omega'}\right]$
$M.1$	$\forall \diamond \in \{\overset{(\cdot)}{\Rrightarrow}, \overset{(\cdot)}{\to}, \dot{\rightsquigarrow}\}. \left[\dfrac{\begin{array}{lll} \Omega \vdash M \diamond (v, T \to U) \dashv \Omega' & \textit{// get pair} & \\ \Omega' \vdash T \dot{\leftsquigarrow} v \dashv \Omega'' & \textit{// calculate right restriction} & \\ \Omega'' \vdash U \dot{\rightsquigarrow} w \dashv \Omega' & \textit{// calculate right restriction} & \\ \Omega' \vdash M \overset{(\cdot)}{\Leftarrow} (m, \_ \to \_) \dashv \Omega''' & \textit{// replace pair} & \textit{if } \diamond = \overset{(\cdot)}{\Rrightarrow} \\ \Omega''' = \Omega' & & \textit{if } \diamond \neq \overset{(\cdot)}{\Rrightarrow} \end{array}}{\Omega \vdash M.1 \diamond w \dashv \Omega'''}\right]$		
$*M$		$\dfrac{\Omega \vdash M \overset{(\cdot)}{\to} \mathsf{borrow}^{\mathsf{s}}\, l\, v}{\Omega \vdash *M \overset{(\cdot)}{\to} v}$	
	$\dfrac{\Omega \vdash M \overset{(\cdot)}{\Rrightarrow} \mathsf{borrow}^{\mathsf{m}}\, l\, v \dashv \Omega' \quad \Omega' \vdash M \overset{(\cdot)}{\Leftarrow} \mathsf{borrow}^{\mathsf{m}}\, l\, \bot \dashv \Omega''}{\Omega \vdash *M \overset{(\cdot)}{\Rrightarrow} v \dashv \Omega''}$	$\dfrac{\Omega \vdash M \overset{(\cdot)}{\to} \mathsf{borrow}^{\mathsf{m}}\, l\, v}{\Omega \vdash *M \overset{(\cdot)}{\to} v}$	

Figure 3.8: readtable

Chapter 4

Evaluation

Ochre has two goals: to allow the programmer to encode strong properties in the type system, and to be efficiently executable on hardware. Section 4.2 evaluates Ochre against the former goal, and Section 4.1 evaluates against the latter.

A secondary goal of this research is to make a pleasant and powerful language with useful abstractions. This is discussed in Section **??**.

The contribution of this research is the design and specification of a language, as opposed to an implementation, so the contributions will be evaluated against how well the design & type checking algorithm lays the foundations for a future implementation.

4.1 Type System

The type system presented must reject well-formed programs, and reject ill-formed ones. This section evaluates whether or not this is the case in two distinct ways: firstly Section 4.1.1 shows the typing rules in action for a collection of programs with should or shouldn't compile. Then, Section 4.1.2 attempts to reason about what properties should and do hold.

4.1.1 Example Type Checks

4.1.2 Properties & Proofs

This section aims to convince the reader that the following property holds:

Which reads "If the abstract interpretation runs to t, and the concrete interpretation runs to v, then $v : t$". It reflects the fact that program execution always starts with a statement in an empty environment.

This soundness property assumes $\dot{\Rightarrow}$ and $\Rightarrow$ instead of only assuming the former and concluding the latter because that would equate to "if it type-checks, it terminates" which is undecidable [Turing, 1937].

The proof is done by induction on the $\dot{\Rightarrow}$ derivation and the $\Rightarrow$ derivation simultaneously. So that a stronger inductive hypothesis can be assumed, we prove this stronger property:

4.2 Performance

maybe explicitly talk about cache coherency and pointer indirection and function pointers

For the presented design to be good, it must describe a language with semantics that can easily be translated into efficient machine code by a compiler. This section discusses the various language features of Ochre and their performance implications.

Due to having a borrow checker, the abstract interpretation introduced in Section 3.3 statically determines when objects are dropped, which means it can insert any necessary memory frees into the resultant binary, removing the requirement for a garbage collector.

Being able to mutate data structures in place also allows programmers to express efficient algorithms provided they don't break the *aliasing xor mutability* invariant, like Rust.

Ochre does not have native machine integers, which restricts the programmer to using Peano arithmetic or similar. This is disastrous for performance, and would not be tolerated in even the slowest languages. However, the design presented, and its type checker are perfectly compatible with integers (they would be similar to atoms), so while this work does not directly include efficient integers, I consider them compatible with it, and adding them would be a matter of engineering. The decision not to include them is discussed further in Section **??**.

As presented, the type system does not support unboxed pairs, which means Ochre programs as currently stated have a lot of unnecessary indirection in their data structures. Much like not supporting machine integers, this is intolerable for a production systems language. However, the type system as presented is compatible with adding them in the future, and adding them at this point would detract from the core concepts. To demonstrate this compatibility, Appendix A.2 lays out a potential method for adding unboxed types.

Ochre does not have efficient contiguous arrays, which hurts the implementation of several dynamic structures, but after unboxed pairs are implemented, contiguous arrays are just many nested pairs. It is unclear how you would efficiently lookup the n^{th} element in such a structure, but I am hopeful it would just be a matter of engineering/adding the right optimizations.

Bibliography

ATS-Home, a. URL https://www.cs.bu.edu/~hwxi/atslangweb/Home.html. pages 12

ATS-Implements, b. URL https://www.cs.bu.edu/~hwxi/atslangweb/Implements.html. pages 12

C gcc vs Classic Fortran - Which programs are fastest? (Benchmarks Game). URL https://benchmarksgame-team.pages.debian.net/benchmarksgame/fastest/gcc-ifc.html. pages 7

The Rust Programming Language - The Rust Programming Language, a. URL https://doc.rust-lang.org/book/title-page.html. pages 19

Rust vs C++ g++ - Which programs are fastest? (Benchmarks Game), b. URL https://benchmarksgame-team.pages.debian.net/benchmarksgame/fastest/rust-gpp.html. pages 7

StackOverflow developer survey, 2021. URL https://insights.stackoverflow.com/survey/2021. pages 2

Are we stack-efficient yet?, November 2022. URL https://web.archive.org/web/20221128082216/https://arewestackefficientyet.com/. pages 7

Thorsten Altenkirch, Nils Anders Danielsson, Andres Löh, and Nicolas Oury. ΠΣ: Dependent types without the sugar. In *Functional and Logic Programming: 10th International Symposium, FLOPS 2010, Sendai, Japan, April 19-21, 2010. Proceedings 10*, pages 40–55. Springer, 2010. pages 16

Karthikeyan Bhargavan, Barry Bond, Antoine Delignat-Lavaud, Cédric Fournet, Chris Hawblitzel, Catalin Hritcu, Samin Ishtiaq, Markulf Kohlweiss, Rustan Leino, Jay Lorch, et al. Everest: Towards a verified, drop-in replacement of HTTPS. In *2nd Summit on Advances in Programming Languages (SNAPL 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. pages 2

Andrew Ferraiuolo, Andrew Baumann, Chris Hawblitzel, and Bryan Parno. Komodo: Using verification to disentangle secure-enclave hardware from software. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 287–305, 2017. pages 2

Son Ho and Jonathan Protzenko. Aeneas: Rust Verification by Functional Translation. *Proceedings of the ACM on Programming Languages*, 6(ICFP):711–741, August 2022. ISSN 2475-1421. doi: 10.1145/3547647. URL http://arxiv.org/abs/2206.07185. pages 11, 14, 20, 35, 37

Graydon Hoare, February 2022. URL https://twitter.com/graydon_pub/status/1492792051657629698. pages 2

Ralf Jung, Jacques-Henri Jourdan, Robbert Krebbers, and Derek Dreyer. RustBelt: Securing the foundations of the Rust programming language. *Proceedings of the ACM on Programming Languages*, 2(POPL):1–34, January 2018. ISSN 2475-1421. doi: 10.1145/3158154. URL https://dl.acm.org/doi/10.1145/3158154. pages 35

Jacob R Lorch, Yixuan Chen, Manos Kapritsos, Bryan Parno, Shaz Qadeer, Upamanyu Sharma, James R Wilcox, and Xueyuan Zhao. Armada: Low-effort verification of high-performance concurrent programs. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 197–210, 2020. pages 2

A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265, 1937. ISSN 1460-244X. doi: 10.1112/plms/s2-42.1.230. URL https://onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230. pages 44

Sebastian Ullrich. Kha/electrolysis, January 2024. URL https://github.com/Kha/electrolysis. pages 20

Leslie Blackett Wilson and Robert G. Clark. *Comparative Programming Languages*. International Computer Sciences Series. Addison-Wesley, Harlow London New York [etc.], 3rd ed edition, 2001. ISBN 978-0-201-71012-0. pages 7

Appendix A

Appendicies

A.1 Formal Verification using (Dependent) Types

The primary motivation behind adding dependent types to a language is so you can perform theorem proving/formal verification in the type system. In some languages, like Lean, this is done to mechanize mathematical proofs to prevent errors and/or shorten the review process; in other languages, like F*, Idris or ATS this is done to allow the programmer to reason about the runtime properties of their programs. However, they are all just pure functional languages with dependent types, whether you choose to use this expressive power for maths or programs the underlying type system is the same.

So the question is how can you represent logical statements as (potentially dependent) types and use the type checker to prove them? This is best understood via a simpler version: proving logical tautologies using Haskell's type system.

Boolean Tautologies in Haskell

The Curry-Howard correspondence states there is an equivalence between the theory of computation, and logic. Specifically: types are analogous to statements, and terms (values) are analogous to proofs. Under this analogy, $5 : \mathbb{N}$ states that 5 is a proof of $\mathbb{N}$.

We can use this to represent logical statements as types. Here is how various constructs in logic translate over to types (given in Haskell).

Logical Statement	Equivalent Haskell Type	Explanation
$\top$	`()`	Proving true is trivial, so unit type.
$\bot$	`!`	There exists no proof of false, so empty type.
$a \Rightarrow b$	`a -> b`	If you have a proof of a, you can use it to construct a proof of b.
$a \wedge b$	`(a, b)`	A proof of a and a proof of b combined into one proof.
$a \vee b$	`Either a b`	This proof was either constructed in the presence of a proof of a or a proof of b.

For example, to prove the logical statement $(a \wedge b) \Rightarrow a$, we must define a Haskell term with type `(a, b) -> a`, which can be done as such:

```
proof :: (a, b) -> a
proof (a, b) = a
```

For another example, we can prove $((a \wedge b) \vee (a \wedge c)) \Rightarrow (a \wedge (b \vee c))$, which you might want to convince yourself of separately before moving on, by providing a Haskell term of type `Either (a, b) (a, c) -> (a, Either b c)`.

```
proof' :: Either (a, b) (a, c) -> (a, Either b c)
proof' (Left (a, b)) = (a, Left b)
proof' (Right (a, c)) = (a, Right c)
```

With this we can construct proofs for logical tautologies, but how do we go further and construct proofs for statements like "If you get any number and double it, you get an even number".

Dependent Types are Quantifiers

Let's now define a function $even$ which returns a type, such that any term of type $even(n)$ is proof that n is even. To do this, $even$ returns a *type*: $\top$ if n is even, $\bot$ otherwise. I.e. $even(4) = \top$ and $even(5) = \bot$. The logical statement $\forall n : \mathbb{Z}.even(2n)$ can be represented by the type $(n : \mathbb{Z}) \rightarrow even(2 * n)$. If we had a term of this type, we could give it any integer n, and it would return proof that $2n$ is even.

This cannot be represented in Haskell, because $(\mathbf{n} : \mathbb{Z}) \rightarrow even(2 * \mathbf{n})$ is a dependent type, hence we need a dependently typed language like Agda. This is an example of Haskell's non-dependent type system not being able to express quantifiers like $\forall$ or $\exists$ over values.

A.2 Supporting Unboxed Pairs

While atoms, functions, and references are unboxed in Ochre, pairs are always heap-allocated. As discussed in Section 4.2, this will hinder the performance of compiled Ochre programs. This appendix lays out a rough plan for adding unboxed types to the formal semantics for Ochre, to make the point that the research as presented is compatible with such an extension.

A potential method of adding unboxed pairs:

1. **Add** $\mathrm{box}\, t$**, a new type which represents an explicit heap allocation**, along with corresponding constructors and eliminators. This is required so the programmer can heap allocate objects whose size is not comptime known.

2. **Edit the abstract interpretations to pass around (type, size) pairs instead of just types**. This would involve all read arrows returning the size of the value being read, and all write arrows taking the size of the data being written.

   Set the size of pairs to the sum of the size of the elements in the pair. Because the type of the right-hand side can depend on the left, this can cause some sizes to be unknown. Because of this, arrows may return an unknown size, and if the user needs to put something of unknown size on the stack, they must put it in a box.

   The size of a match statement is the largest size of any of its branches.