Programs and Proofs in Practice

A Grounded Theory on How People Use Dependently Typed Interactive Theorem Provers

RUBEN BACKX, Delft University of Technology, The Netherlands
SÁRA JUHOŠOVÁ, Delft University of Technology, The Netherlands
WOUTER SWIERSTRA, Utrecht University, The Netherlands
JESPER COCKX, Delft University of Technology, The Netherlands

Dependently typed interactive theorem provers (DTITPs) are powerful tools that help improve software correctness. However, they currently serve niche communities, and have not yet been adopted in mainstream software development processes. Through this work, we aim to understand *how people use dependently typed interactive theorem provers* to provide a solid, layered understanding of the starting point for any future advancements in DTITP design. We apply the *small-scale*, *emergent-mode* Socio-Technical Grounded Theory methodology to develop a theory grounded in data from interviews conducted with DTITP users with varying levels of experience, using Agda, Rocq, and Lean. Our results show how DTITP programmers use types to plan and constrain their programs; make trade-offs in how they represent data; and continuously interact with the compiler to construct a program iteratively. Based on our findings, we have formulated five recommendations for the developers and maintainers of DTITPs who wish to make their work more accessible to a broader audience of software developers.

Additional Key Words and Phrases: interactive theorem provers, dependent types, grounded theory

1 INTRODUCTION

Dependently typed interactive theorem provers (DTITPs) such as Agda [42], Rocq [51], and Lean [16] provide a unified system for programming and formal verification, and have gained traction within the academic community. It is not uncommon for papers on programming languages to be accompanied by some form of mechanised proof; benchmarks such as the POPLMark challenge [5] have helped push the further development of these systems. Lean has gathered momentum among mathematicians. Mathlib, a community-built library of formalised mathematics, now contains mechanised proofs of tens of thousands of theorems, formalising both a substantial part of undergraduate mathematics and recent research results. The CompCert project uses Rocq to verify a real-world C compiler [34]. Yet despite all these successful academic applications of DTITPs, their impact on day-to-day software engineering remains much more limited.

In principle, these systems are prime candidates for implementing safety-critical software, where failure has severe consequences. In practice, however, DTITPs have not (yet) been widely adopted by industry. Instead, most software is verified using methods that are more integrated into mainstream development processes, such as code review or automatic software testing. The shared understanding of the community is that the DTITP learning curve is very steep, and that the cost of using DTITPs for verification is too high for day-to-day industry use. Additionally, previous research has shown that new users find it hard to imagine writing software that end users can interact with in a DTITP [30, p. 166].

How can we make DTITP technology more accessible to a wider range of programmers? We set our aims high: we are not only interested in incrementally improving the current generation of tools, but also in understanding whether the design philosophy behind DTITPs and their ecosystems matches the needs of the wider software development community. To do so, we need to understand our starting point: we need to understand how people *currently* use dependently typed interactive theorem provers. A good understanding of the current state of affairs serves as the starting point for any direction, whether that be our goal of more widespread adoption or further tailoring a DTITP towards a particular user group. This leads us to the following research question:

RQ: How do people currently use dependently typed interactive theorem provers?

We answer this research question by applying Socio-Technical Grounded Theory (STGT) [27, 28] to twelve interviews conducted with DTITP users from different backgrounds, working in Agda, Rocq, or Lean. STGT is a particularly good methodology for answering "how" questions and is best suited to "studying topics that are socio-technical in nature and where the research team wishes to produce rich and layered qualitative findings" [28, p. 45]. Such rich, layered findings allow us to provide a full theory on how people currently use dependently typed interactive theorem provers, which in turn allows us to provide data-grounded recommendations for making DTITPs more accessible.

Our results present a theory consisting of seven hypotheses, explaining how DTITP programmers use types to plan and constrain their programs, make trade-offs in design decisions, and interact with and respond to the compiler[1]. We use these hypotheses to provide five recommendations for DTITP maintainers and developers to help them make their work accessible to a broader audience, and discuss the work that has already made progress in these directions in Section 6. By grounding our recommendations in the collected data, we provide a user-oriented basis for future work in DTITP usability. We view this work as a stepping stone towards bringing dependently typed interactive theorem proving into mainstream development processes.

2 METHODOLOGY

To answer our research question, we employed Socio-Technical Grounded Theory [28], a methodology for forming new theories grounded in (usually qualitative) data. In this *small-scale*, *emergent-mode* version of such a study, we conducted *online* think-aloud programming sessions and semi-structured interviews with twelve DTITP users. We analysed the data using our *constructivist* research paradigm as a team of researchers focusing on dependent type theory, functional programming, and the usability of advanced type systems. The resulting theory is presented in Section 3 as a set of hypotheses supported by quotes from the interviews, and we provide recommendations for DTITP maintainers based on our findings in Section 4.

2.1 Socio-Technical Grounded Theory

Socio-Technical Grounded Theory (STGT) is a version of the Grounded Theory methodology [10, 15, 21] defined by Hoda [27, 28] for projects that study *socio-technical* phenomena. It consists of a basic and advanced stage, producing preliminary hypotheses and a mature theory respectively. Grounded Theory is applied in iterations, where each iteration consists of the research team (1) collecting new data, (2) analysing the new data using *coding* and *constant comparison* to organise it into concepts and categories, and (3) deciding where to collect data next. The process stops when an iteration does not provide any new insights, i.e., upon reaching *theoretical saturation* [21, p. 61]. Throughout the entire process, the team writes memos — notes of "ideas about codes and their relationships as they strike the analyst" [20] — and use them to discuss and form the theory.

Coding is "the process of closely inspecting, deeply making sense of, and inferring meaning from data and giving those meanings some labels or names" [28, p. 232]. During the basic stage, we used *open coding*, labelling the data "inductively and comprehensively, with an open mindset, to enable emergence of information and insights, without looking to find anything specific in the data" [28, p. 233]. For example, the following interview snippet was coded as "`difficulty refactoring`":

> [P3] Agda is very good at this exploratory stuff, but when it comes to maintenance and refactoring…

[1]By "compiler" we mean the tool that parses, scope checks, type checks, and compiles or interprets the code.

round	id	identified as	gender	recruited via	think-aloud	editor
R1	P1	Formal Methods Engineer	Male	Personal invite	Sudoku	Emacs
Agda	P2	Researcher (type theory)	Male	Personal invite	Sudoku	Emacs
	P3	Researcher (type theory)	Male	Personal invite	Sudoku	Emacs
R2	P4	Student (BSc Computer Science)	Male	Personal invite	Sudoku	VS Code
Agda	P5	Student (BSc Computer Science)	Male	Personal invite	Sudoku	VS Code
	P6	Student (BSc Computer Science)	Female	Personal invite	Sudoku	VS Code
	P7	Student (BSc Computer Science)	Female	Personal invite	Sudoku	VS Code
R3	P8	Software Developer	Male	Personal invite	own work	VS Code
Lean	P9	Researcher (mathematics)	Male	Online sign-up	own work	VS Code
	P10	Researcher (mathematics)	Male	Online sign-up	own work	VS Code
R4	P11	Researcher (type theory)	Male	Personal invite	own work	VIM
Rocq	P12	Researcher (type theory)	Male	Online sign-up	own work	VS Code

Table 1. An overview of the participants, how they were recruited, and what they worked on during the think-aloud part of the interview (Sudoku = a prepared assignment on validating an insertion into a Sudoku; VS Code = Visual Studio Code)

Once we started seeing emergent categories and hypotheses, we moved on to *targeted data collection* [28, p. 299], using *targeted coding* to focus on only the most significant / promising concepts and categories. The process was still inductive, but fresh codes were constantly compared with other codes, and similar codes were grouped into concepts. For example, during a subsequent interview, we added the following quote into the "`refactoring`" concept:

> [P7] But I just tried to make things as simple as possible and build up on top because it's really inflexible to work with it.

This concept was later grouped into the "`ecosystem`" category, along with concepts such as "`standard library`", "`documentation`", and "`error messages`". The concepts, codes, and quotes formed the basis for our hypotheses and recommendations, presented in Sections 3 and 4.

Because it was manageable for a single analyst to code the dataset, the whole coding process was conducted by the first author to ensure consistency. However, as suggested by Hoda [28, p. 260], the codes were regularly reviewed by the other researchers on the team, and the resulting concepts and categories were discussed in weekly meetings. We acknowledge the creative subjectivity of our codebook (provided in the data repository [4]), and declare our *constructivist* research paradigm [28, p. 46]. We discuss how this approach might have affected our results in Section 5.

2.2 The Interviews

We collected our data in four iterations of *online* interviews with a total of twelve participants. Table 1 contains an overview of their background, the DTITP and editor they used, and how they were recruited. The interviews consisted of an hour-long think-aloud session during which they used the DTITP and a subsequent 30-minute *semi-structured* interview [28, pp. 177–188]. They were conducted over Zoom, and the recordings stored on GDPR-compliant servers. Participants were asked to sign an informed consent form, agreeing to have their voice and screen recorded, and the anonymised transcripts of their interviews published under an MIT Licence [4].

The interview recordings were transcribed using Whisper [45], diarised using Whisper X (run locally), and then checked and corrected manually. We did not transcribe the think-aloud sessions, but instead wrote down whenever the participant did anything interesting, and then asked about it

during the interview. This meant that the think-aloud data was analysed as part of the interview, and the original recording was referenced if additional details were needed. After a transcript was fully processed, we rewatched the full recording and wrote down and coded anything we missed.

Pilot. The first two interviews were conducted as pilots with participants who were experienced Agda users. We asked them to have their own installation of Agda ready and to work in the environment they always do when using the DTITP. For the think-aloud session, we gave them 45 minutes to implement the following assignment:

> Given a valid, partially-solved Sudoku and the next move done by a player (putting a number somewhere), determine whether the resulting Sudoku is still valid. In classic Sudoku, the objective is to fill a 9×9 grid with digits so that each column, each row, and each of the nine 3×3 subgrids that compose the grid (also called "boxes", "blocks", or "regions") contains all of the digits from 1 to 9[2].

We did not elaborate any further, with the aim to let the participants approach the assignment however they saw fit. The participants were asked to share their screen and to voice all their thoughts as they were programming. We prepared a rough outline for the subsequent 45-minute interview, focusing on the challenges in the assignment, the different use cases of Agda, the features they like and do not like, and the setup they use when programming in Agda.

We found that the recording setup worked well, and that we were gathering the type of information we were looking for. However, we did notice that the participants found the allocated 45 minutes too short for the think-aloud session, and adjusted the schedule to 60-minute think-aloud sessions and 30-minute interviews. The outline of the interview stayed roughly the same, but we decided to drop discussions on their programming setup, since we could derive those ourselves from the recordings. We did not ask each participant every question and preferred specific questions relating to the think-aloud part of the interview. The interview protocol and question template is available in our data repository [4].

Round 1: Experienced Agda Users. Since the pilot interviews went well, we used them as data in the first round, and conducted one more interview to round up the data collection of the first STGT iteration. This meant that the first set of participants were experienced Agda users, all recruited via personal invite. After using the basic stage of STGT to analyse the data collected from this first round, we already saw a variety of hypotheses emerging with two avenues of interest for further data collection:

(1) the participants struggled with completing the Sudoku assignment we had given them even though we thought it would be relatively simple; and
(2) the participants seemed to have an aversion to using the standard library for any part of their solution.

Since the concepts and categories were already forming relatively strong emerging hypotheses, we moved on to the advanced stage of STGT and targetted participants who could fill in the gaps.

Round 2: New Agda Users. The participants from the first round had indicated that they spent a large portion of their time planning their program, leaving them with less time to actually focus on solving the assignment. We suspected that this was because they were already thinking about the design decisions that would make verification easier in the long run, and decided to explore that hypothesis by interviewing users who were new to DTITPs. We invited a set of four computer science students who had just had a two-week introduction to Agda at our university. Notably, these new users were actually able to get *further* in the Sudoku assignment than the professional

[2]This definition was copied from the Wikipedia page.

participants, but they did so by using almost none of Agda's dependently typed features. We also noticed that when the students *did* try to use a dependently typed feature, they struggled more than the experienced users and required more planning in those scenarios.

Round 3: Mathematicians Using Lean. To explore the aversion to using the standard library in Agda, we decided to invite three users of Lean, a more recently developed DTITP. We had heard that Lean has a good ecosystem, active community engagement, and a popular library for mathematics called Mathlib. Unlike Agda, proving in Lean is usually done using tactics, which meant that our concepts and categories expanded to account for this different proving approach. Additionally, many of Lean's users are mathematicians looking to formalise their theorems, and we decided to include them in this round of interviews, investigating how users who are not necessarily computer scientists interact with a DTITP. At this point, we also evaluated that there was no more to learn from the Sudoku assignment, and we asked each participant to work on their own projects during their think-aloud sessions, allowing us to observe how a DTITP is applied in real-world scenarios. We reached out to interested participants with the relevant experience who had filled in an online sign-up survey distributed over our social media and via DTITP community channels.

Round 4: Rocq Users. For the fourth round of interviews, we invited two Rocq users. This was done to verify that our hypotheses are robust and extend to a wider range of DTITPs. One participant was invited via personal invite and one was invited via the online sign-up. In this final iteration, we made minimal adjustments to our categories, and were able to form our mature theory.

3 RESULTS

Based on the four rounds of interview conducted with a total of twelve participants, we were able to form seven overarching hypotheses about how people use dependently typed interactive theorem provers. These are related to how DTITP programmers plan and design their programs, how they interact with the DTITP, and the control they wish to have over certain features. The anonymised transcripts of the interviews are available in our public data repository under an MIT Licence [4].

3.1 Planning

H1: DTITP users use types to create a plan for their program and to constrain the solution space. They adjust this plan as necessary throughout the process.

During the first round of interviews, we gave participants what we assumed to be a relatively simple programming assignment, asking them to validate an insertion into an already-valid Sudoku. Surprisingly, we found that most expert participants started over-engineering the solutions and did not get as far as we expected. This lead us to give the same assignment to the participants from the second round, who were new to using a DTITP, but not new to programming. As expected, they planned less and got further within the same time frame, but remarked that they could have made more progress using a programming language with which they were more familiar. The expert-level participants explained that they had learned to carefully plan their programs before they start writing any code, in order to choose a definition that will not require major refactoring later. The students seemed to be aware of this notion too, remarking that they needed to think further ahead than in other programming languages.

> [P7] If this were any other language, I would just type some stuff. It wouldn't work but, immediately, I would be able to fix it. But here, I need to think ten steps in advance.

Participants planned their solutions by sketching outlines for their programs with data types and type signatures, and then revised and refined them throughout the entire think-aloud session.

They claimed that thinking far ahead is necessary when programming in a DTITP, as the choice of the "plan" — and therefore the choice of the (data) types — can determine the rest of the program. The students we interviewed considered this tight coupling between types and implementation to be a downside, but expert participants overwhelmingly saw this as a benefit.

Types constrain the solution space for the implementation which, in some cases, can allow for the only possible solution to be automatically filled in. This moves the cognitive load for the programmer from the implementation to the planning phase. This process is often facilitated in DTITP by "holes"[3], which act as placeholders for not-yet-implemented parts of the code. These holes carry type information about the goal as well as the available context; developers use them to mark unfinished parts of the program.

> [P2] Holes are super useful. Also, when I'm programming Haskell, I program in the same style now. [...] Because, typically, the types tell you where to go.

The type information can be leveraged by adopting some style of "hole-oriented development", which many participants did. In this style of development, the programmer scans through the program and looks for the hole they want to fill based on the local context and goal. They then attempt to fill the hole or, if that proves too difficult, fill the hole *partially*, i.e., by filling it with an expression containing more holes. Alternatively, the programmer can revise the types to make it easier or even possible to fill the hole. Revision can be the only viable option if the programmer's original plan turns out to be flawed — which happened in almost every think-aloud session.

Participants told us that this process of filling holes and revising the types feels like solving a puzzle, which makes using a DTITP not only useful, but also fun. They found formalising a proof similar to playing a game, where strategy plays a big role in determining how smooth the process is going to be. If they eventually find out that their strategy was not designed well, they have to start refactoring their code — a point in which the process can stop being fun.

> [P3] Okay, I want to change something in my data type. Maybe I want to, you know, add a field to a constructor or remove a field from a constructor or add a constructor to a data type, and then you end up having to do all this manual work, which is too bad.

Refactoring, they explained, is typically not an enjoyable task in a DTITP. If a type signature needs to be adjusted (e.g., the type of an argument needs changing or a constructor has to be added to a data type), the programmer has to go through the entire code base and manually fix all the places where that signature is used. According to our participants, the current tooling does not help with automating such a scenario. They consider this to be an oversight.

3.2 Design

When planning their programs and proofs, our participants considered the following design choices:

- whether to encode the invariants within their data types (intrinsic verification), or to prove them separately (extrinsic verification);
- the degree to which their definitions should be generalisable (abstraction); and
- the style in which they write their code and their proofs (code style and convention).

H2: DTITP users aim for a "sweet spot" between intrinsic and extrinsic verification, influenced by the design of the DTITP they are using.

An approach where properties are represented and verified extrinsically offers simple data types, but decoupling the properties from the definitions has its drawbacks. For example, establishing

[3] A hole is denoted using a question mark (?) in Agda and `sorry` in Lean.

certain invariants may rely on additional properties that need to be proven and invoked separately. As a result, programmers need to know where to find the auxiliary properties their proof requires, which can be challenging if there are many places to look. More frustratingly, functions may still require the programmer to implement "bogus" cases for the sake of totality, even though the extrinsic property clearly shows those cases to be impossible.

When taking a more intrinsic approach, where the proofs are stored directly in the data types, the programmer can easily find the proofs they need and does not need to provide implementations for impossible program branches. However, by requiring all instances of a data type to contain a proof, the possible variants that can be created are constrained. This can be beneficial, but can also get the programmer stuck — especially when re-establishing the intrinsic proofs is non-trivial. If it turns out that the invariant is too strict for a certain scenario or needs to be temporarily broken, the code needs to be refactored to accommodate for this. This point is illustrated by what some participants did with the Sudoku assignment. Some experienced participants first designed their solution in an intrinsic manner, producing something akin to the following:

```
record Sudoku : Set where
  field
    grid : Vec (Vec (Maybe (Fin 9)) 9) 9
    noDuplicatesInRows : (i : Fin 9)
      → lookup (noDuplicates (rows grid)) i (≡) true
    noDuplicatesInColumns : (i : Fin 9)
      → lookup (noDuplicates (columns grid)) i (≡) true
    noDuplicatesInBoxes : (i : Fin 9)
      → lookup (noDuplicates (boxes grid)) i (≡) true
```

This definition encodes the "no duplication" properties directly into the structure holding the data that represents the current state of the Sudoku. The participants would attempt to solve the assignment by first applying a move to the Sudoku, and only then checking whether the Sudoku is still valid. They quickly came across the problem: applying a potentially invalid move to a Sudoku is not possible using a representation that allows only valid Sudokus to be constructed. After realising this, the participants moved towards a more extrinsic approach:

```
record Sudoku : Set where
  field
    grid : Vec (Vec (Maybe (Fin 9)) 9) 9

valid : Sudoku → Bool
valid sudoku = all noDuplicates (rows sudoku)
            && all noDuplicates (columns sudoku)
            && all noDuplicates (boxes sudoku)
```

This representation is what most participants ended up with.

[P1] Probably you could encode valid Sudokus in Agda somehow. It might also be a very ugly and unwieldy representation. That's one end of the spectrum and the other end of the spectrum is just having a bunch of lists. And I think there's something in between there.

Where exactly the ideal solution lies on the spectrum depends on the DTITP. Generally, a DTITP that enables easily searching for proofs, such as Lean, is going to lean more towards the extrinsic side of the spectrum, whereas a DTITP that is designed more around embedding the right properties in the data, like Agda, leans more towards the intrinsic side. Some participants mentioned the existence of a certain "sweet spot" on the spectrum of representing invariants, specific to the problem at hand and the DTITP used to solve it. They were looking for such a sweet spot multiple times during the think-aloud sessions.

H3: DTITP users search for the "right level" of abstraction in their code.

Across all the DTITPs, programmers indicated that there is a certain level of abstraction they aim for when designing their data types and proofs, differing per task and DTITP. Participants explained that when contributing to a library, whether intended for personal use or to be shared with the community, they often aim for a higher level of abstraction to ensure the proofs can be applied in a wider range of contexts. While many participants did not start out with the goal of contributing to a library, they often realised the proof or function they were working on might be useful in the future, leading them to refactor their work so that it would be more abstract. For example, one participant started out with the following definition for a function that splits a vector into equal parts:

```
chunk : {A : Set} → Vec A 9 → Vec (Vec A 3) 3
```

They later refactored it into a more generally applicable definition:

```
chunk : {A : Set} → {n m : Nat} → Vec A (n * m)
      → Vec (Vec A m) n
```

When working towards their main goal, participants usually would not alter the level of abstraction, at least until they were finished. Once finished, they sometimes started abstracting parts of their work in order to be able to reuse it later. When *defining* the main goal or a sub-goal, however, a lot of thought was put into determining the right level of abstraction.

[I] "Do you think that the generalisation actually gives you better insight into the problem?"

[P1] "Yes [...] it's also going to be a shorter solution if you do it more generally."

[I] "Do you ever find yourself reusing your old code that you've made more general?"

[P1] "Yes."

It should be noted that the Sudoku assignment description did not require the solution to support any size other than 9×9. Even if a participant chose to hardcode some part of their solution, they often made some other part more abstract than it needed to be, e.g., by having some sort of `rows` or `columns` function work for any size of list. In later rounds of interviews, where participants were working on their own work, they also made some parts of their work more abstract than it needed to be.

[P3] And then I did this really icky thing, which is I knew it's like three, so I could just pattern match on it and that was very unsatisfactory for me because it's like, oh, I'm hacking my way through here.

Despite this preference towards abstraction, most participants also seemed to have a notion of a solution that would be "too abstract". Participants considered a lemma or function too abstract when it was no longer recognisable as a solution to the specific problem it was intended to solve.

Both the field in which a participant worked and the DTITP they used contributed to their understanding of the "right" level of abstraction. In mathematical disciplines, participants followed the conventions of the Mathlib lemmata defined for that discipline. More programming-oriented tasks that did not require as many proofs were allowed to be more concrete. The most abstract definitions were found in type-theoretical work, where participants often leveraged the type system of DTITPs to their full extent. Most Agda programmers told us they have a personal library of definitions, from which they copy-paste as necessary. Any definition they considered to be a nice addition to their library would be written using a higher level of abstraction.

H4: DTITP users follow a code style geared towards maintainability and community conventions.

Besides ending up with a compiling program, our participants cared about a secondary goal: their code looking "nice". What exactly "nice" meant to each participant differed slightly and was influenced by the DTITP they were using.

> [P10] So, at the end, of course, you want a proof that works, because you want a proof, and there's no proof that doesn't work, by definition of proof. But you want a nice proof, you want an elegant proof. I mean, just having a proof is the bare minimum.

In some DTITPs, a proof was considered "nice" if it was short, but not too short. We saw this among the users of the more tactic-oriented DTITPs, i.e., Lean and Rocq. What constitutes "too short", participants said, are proofs that use tactics and functions so complicated, they are made unintelligible. For Agda users, there was no such thing as a proof that was "too short" — the most prominent opinion seemed to be "the shorter, the better".

Our participants explained their preferences were mostly guided by legibility and maintainability. In Agda, a shorter proof is often more elegant since language elements map roughly one-to-one to reasoning steps. However, since tactics can perform arbitrarily complex transformations of the proof state, a short proof in a DTITP with tactics offloads much of the reasoning to the tactic implementation. If the tactics involved are complex, e.g., performing automated proof search, it can be hard to comprehend the structure of the proof.

> [P10] What I would call a nice proof, at least a nice Lean proof, is a proof that is at the same time very short, but contains all the relevant information to understand how the argument went. So if it becomes too short, so it hides some ideas or some insight [...], then it's not a nice proof.

Another style consideration for most participants was whether to use external libraries. Most participants did not want to use unofficial libraries, even if writing their own implementations required more work. This was usually because the libraries were difficult to install or because they did not use the same code style as the participant. One of our participants spent their think-aloud rewriting an older library in a "more modern" style. There was, however, a division between the Agda users and the rest on whether to use the standard library[4]. Lean and Rocq users expect others to be familiar with the standard library, so using it makes their own code more understandable. Agda's standard library, on the other hand, is not as widely used, and not all Agda programmers are generally familiar with its content or code style. As a result, Agda users prefer putting a custom definition in the same file, keeping the development more self-contained and easy to understand.

[4]Mathlib is so commonly used in Lean that we consider it to be a "standard" library for the purposes of this paper.

Finally, the tactic-oriented DTITPs had an additional code style consideration which dictates that "powerful" tactics — that is, tactics with a high degree of automation — should not be used in the middle of a proof. Such tactics might be able to do more than solve the current sub-goal in the proof; if any later refactoring changes the sub-goal at that point, the tactic might change the proof state in unexpected ways, potentially causing an unexpected error later down the proof. Participants who *did* use a powerful tactic in their first implementation (such as `simp` in Lean) later replaced it with a restricted version that would do only exactly what was expected (such as `simp only` or `apply`). Powerful tactics at the end of a proof were not considered a problem, as they cannot affect any later lines of the proof.

3.3 Interactivity

Several participants used the term "conversation" to refer to their interactions with the DTITP.

> [P2] I feel like working with Agda is like having a conversation with a partner, right?

H5: DTITP users write code by conversing with the compiler, which can act as an oracle, but also request more information from the user.

With the combination of restrictive types and interactive holes, programming in a DTITP can often feel like conducting a conversation with the compiler. This interactivity begins with a (partial) type definition and a hole which can help guide the implementation process and break it down into smaller steps. A series of such steps then guides the programmer to a complete solution, or signals when further progress is impossible and something needs to be adjusted.

Using holes, the programmer can ask the compiler for local type information ("What is the type of this variable?"), request real-time feedback on potential solutions ("Can I apply this tactic here?"), or even delegate the solution search to the compiler ("Can you split this variable into cases?"). Our participants knew that the system knows more about the exact types in their program than they do. They frequently used this to their advantage by, for example, using a keyboard shortcut to explicitly ask the system for a target type or by referring to the window displaying components currently available in the context.

> [P1] I rely a lot on the type checker if I write Agda.

Conversely, the compiler sometimes requests more input from the programmer to complete the program, usually by providing an error message. For example, if the programmer does not apply a function to the correct number of arguments, the Agda compiler would report "`Error: [UnequalTerms] X → Y !=< Y when checking that the inferred type of an application matches the expected type`", as a prompt for the programmer to supply more arguments. In more obscure error messages, the system can simply report "`Error: unsolved meta variable`" followed by a list of generated unification constraints, essentially telling the user it cannot automatically infer the correct type without more information from the user. Similarly, the following Lean snippet produces "`don't know how to synthesize implicit argument 'n' @Vector.mk Nat ?m.29 { toList := [1, 2, 3] }`":

```
let v := Vector.mk (Array.mk [1, 2, 3])
```

What the compiler means to say is "you need to provide a value for the second argument of the `Vector` constructor". Adding `rfl` as the second parameter makes the error go away.

H6: More experienced users are able to better respond to the DTITP's requests.

> [P8] As a user of Lean, I'm not bothered. I find error messages good enough. I mean, for example, the error message about universe is not matching. I wasn't looking closely at it, but it gave me the information [that] something [was] wrong with levels. That was fine.

Our experienced participants could, in most cases, "translate" error messages into concrete requests for information, but the beginners had a lot more trouble interpreting them. They understood the proof assistant wanted *something*, but had difficulty pinpointing the problematic part of the code.

> [P4] I was using AI tools basically to explain to me the errors because I don't understand the errors.

In some IDEs, system feedback is real-time, providing feedback as soon as the user finishes typing. Our participants were divided on whether this was a good thing. If someone was working in an on-demand IDE (such as Emacs with keyboard shortcuts), they also preferred on-demand feedback, whereas participants working in real-time IDEs (such as VS Code with continuous checking) preferred real-time feedback. Participants who preferred on-demand feedback criticised real-time feedback for its tendency to interrupt them while they were still thinking. On-demand feedback proponents argued that being interrupted helped them stop and think before going down a potentially infeasible path. Some participants mentioned they would even prefer a feature where the system would try to falsify their claims automatically, warning them before they would attempt to prove a false statement. These participants had previously worked with Isabelle [41], an interactive theorem prover which has such a feature but is not based on dependent types.

> [P8] I think they'll get that eventually, Isabelle's ability to tell you that a theorem is unprovable using something like QuickCheck automatically. And the automatic part is interesting. Like this kind of proof assistant point of view, like I'm really trying to help you as a user as you go. And if you make a typo in a theorem statement, you know, like two variables that should be the same or different names, and it's just obviously not satisfiable, then just tell me before I even start coding.

3.4 Control

H7: DTITP users want the entire system to behave more like they think.

In general, the participants were positive about the experience DTITPs provide. One common source of frustration, however, was having to "over-explain" reasoning steps. Frequently, the programmer could clearly see the steps ahead that would get them to their goal, but the DTITP did not. When this happens, the programmer has to manually type out the steps, which feels tedious.

> [P3] But then at some point I got this problem with [data types]. I was using pattern matching on a let bound thing, which was not a record, like, oh, man, there's only one constructor. I can either change it to a record or, you know, fiddle around a little bit with stuff. And that was a distraction in the development, I feel.

One such source of frustration is related to the unfolding of definitions. When a function is applied to known arguments, the resulting term can potentially be evaluated further by inlining the function definition and β-reduction. This can be useful, potentially leading to simpler goals, but can also make it harder to recognise the remaining sub-goal or identify which lemmas can be

used to complete the proof. Several participants expressed their wish to have control over when and where unfolding is applied.

> [P2] […] you really want to be careful about where you unfold and how you unfold things. And sometimes it's really useful to be able to see the non-normalised goals. Sometimes it's really useful to be able to see the fully normalised goal.

4 RECOMMENDATIONS

During our conversations with the participants, we noticed that there was some kind of consensus about certain DTITPs being good for certain niche tasks. Multiple participants stated that their choice of DTITP for a particular project is highly dependent on the nature of that project. For projects involving complex maths, Lean is the most obvious choice, given the existence of MathLib and custom proof automation. Exploring type theory or programming with dependent types is easier in Agda. The creation of a verified program, such as a compiler or a piece of embedded software, is more often done in Rocq. It is important to emphasise that, theoretically, the DTITPs are interchangeable, but depending on the task at hand, one DTITP might provide a smoother experience than another.

Additionally, most participants indicated they do not see much use in using their DTITP of choice outside their particular niche.

> [P1] […] users that do use it, I think, like it a lot because it's for its specific purpose, and that's sort of a programming environment where you can work with dependent types.

In this section, we therefore discuss recommendations on how to make DTITPs accessible for a wider range of programmers in more mainstream contexts. We do not think that all DTITPs should strive towards this goal, since we saw that our participants valued how these DTITPs currently apply to their niche. Changes that would make a DTITP more appealing to a wider audience may also make it less appealing to its current user base. However, we believe there is value in making DTITP technology more accessible, and provide recommendations grounded in our data on what to focus on. In Section 6, we discuss work that has already made progress in this direction.

R1: Keep it simple and design for inexperienced users first.

Learning to use a new programming language is not easy, and we identified several areas where we observed learners and experts struggling. Most of these obstacles are not unique to DTITPs, but the addition of dependent types both magnifies the obstacles for the user and adds an extra layer of complexity to the potential solutions. We zoom in on a few design considerations that were brought up during the interviews, but our overarching recommendation is to prioritise designing for new rather than expert users.

Installation. In preparation for the interviews, some participants had to reinstall their DTITP, which was not always a smooth process and many participants commented on it at the start of their interview[5]. The Agda installation documentation, at the time of writing, assumes the reader understands how to put a binary on their path. Furthermore, the editor and standard library have to be installed separately. This lead to some participants having to conduct their interview without standard library support, much to their frustration. Rocq and Lean have comparatively straightforward installation instructions, but troubleshooting can be difficult for all of them. Participants would have liked a more straightforward installation process, preferably directly from an editor they are already familiar with.

[5]Some of these comments are not caught on our recordings, because they were usually made directly upon joining the call.

Design of Standard Libraries. There is some tension in the design of standard libraries. On the one hand, they typically aim to be widely applicable and reusable but, often, the most general definition is not the easiest to understand. Consider, for example, the type signature of function composition in Agda's standard library:

```
∀ {A : Set a} {B : A → Set b} {C : {x : A} → B x → Set c}
    → (∀ {x} (y : B x) → C y) → (g : (x : A) → B x)
    → ((x : A) → C (g x))
```

This type is vastly more general than the corresponding simply typed version:

```
∀ {A B C : Set} → (A → B) → (B → C) → (A → C)
```

Not only does the dependently typed version abstract over the universe levels involved, the functions being composed may themselves be dependently typed. Yet this type is hardly recognisable as function composition! Similarly, many properties abstract over the notion of equality being used, e.g., when parametrising over a setoid in algebraic structures such as monoids or rings. During the interviews, participants complained about the effort needed to specialise a library function for their use case.

> [P2] I like the standard library so abstract because that means that it is applicable to many things, but it is also really painful sometimes too. Because everything is so abstract, you really have to put in work to apply [the functions] there to get the functionality that you want.

Participants programming in Lean, on the other hand, did not face these issues. One possible explanation we offer is that many of Lean's libraries use type classes, whose instances are automatically synthesised when necessary. This leads us to believe that the adoption of highly general definitions in a standard library requires targeted language support to be employed effectively.

Along a similar vein, the standard libraries are difficult to search through due to their architecture as well as the difficulty of recognising general type signatures as something applicable to the programmer's current context.

> [P3] I want to find the lemma that plus is commutative in the standard library. It's somewhere hidden in the fact that blah, blah, blah, or whatever. There's like seven layers of records and padding around it where I think, oh, I shouldn't.

Often, the only method of finding the definitions programmers are looking for is to search for the expected name of the lemma in an entire directory, which can be difficult when the naming convention is not consistent. Some libraries, such as Mathlib and the Lean standard library, offer a web page with search functionality, and there exist tactics that can search for all applicable proofs in the current context. Still, even participants working in DTITPs with these search options would have liked better searching to be available directly in the IDE. Additionally, two participants expressed their desire for a search by "lemma shape", a template of the theorem they are interested in (e.g., `IsEven _ -> divides 2 _` for "I am looking for some implication which can tell me something is divisible by two if I give it an even number"). We conclude that good in-IDE searching would contribute to making the entire system easier to use.

Syntax. Juhošová et al. [30] identifies *syntax* as a learning obstacle of Agda. In the end, however, only the student participants complained about syntax issues and Unicode characters. This confirms that it is indeed a learning obstacle, but perhaps an obstacle that can be overcome.

> [P6] I think it's because [the syntax is] very, I don't want to say strict, but very specific. Even when writing this, because I forgot a space, then it was not compiling.

Nonetheless, this obstacle to learning should be taken seriously. The only participants who expressed an overall negative sentiment towards DTITPs were the students; we deem it unlikely they ever willingly choose to use a DTITP again. Furthermore, there was consensus between some expert and student participants: the usage of Unicode characters and specific notation is fine in contexts where their meaning is universally understood. Outside those contexts, however, highly specific custom notation obscures the intended meaning of proofs and programs. We therefore believe that syntax does not matter, except when it gets in the way of the learning experience, and effort should be made to report such syntax errors as clearly as possible.

Error Messages. Both students and experts complained about unclear error messages during their think-aloud sessions. Even experts were not always able to correctly interpret the error message and identify the relevant issue in their code. Furthermore, error messages are sometimes phrased using terms that are unfamiliar to beginners, for example, referring to meta-variables or unification errors. Even though we saw that it is possible for users to learn the compiler's jargon, these systems could be made more accessible by providing error messages with fewer technical terms, focusing on how to resolve the problem rather than the internal compiler error.

Working with Dependent Types. Programming and proving with dependent types sometimes exposes details about the underlying type theory, to which our participants directly attributed a class of problems we dubbed "representation problems". These representation problems share a common cause: the choice of data representation and function implementation has a significant impact on the development and proof effort — much more so than when doing pen and paper proofs. This problem was most prevalent among the mathematicians who were not familiar with dependent type theory:

> [P9] It uses dependent type theory and I'm a mathematician. I'm not that familiar with dependent type theory, and it's just things that are really obvious to me are non-trivial because the types are different.

Types in DTITPs are considered equal up to some notion of *conversion* that includes at least β-equality. As a result, different implementations of the same function may lead to different proofs. To illustrate this point, consider the usual definition of addition on natural numbers by induction over the first argument. The proofs that $0 + n = n$ and $n + 0 = n$ hold for all n are vastly different: the first holds trivially by evaluation; the second requires an inductive argument. Unsurprisingly, proofs involving *equality* were one of the greatest sources of such representation problems. Agda, Rocq, and Lean all have several notions of equality (including at least definitional equality, propositional equality, and setoid equality), and it is not always possible to convert between them.

Another common complaint we observed was related to how functionality is bundled. In many DTITPs, there is more than one way to do so: records and instance arguments in Agda; structures and type classes in Rocq and Lean. As a result, we observed frustration among the participants when combining different APIs that use different language features to model conceptually similar notions.

Although representation problems seem inherent to DTITPs, we argue that it is worth exploring how they can be mitigated or at least minimised. The goal here is not to remove any trace of dependent types, since many participants found their guidance useful. On the contrary, we expect that the rich static type information that dependent types provide can be leveraged to help guide users.

R2: Prefer existing, "good" solutions over theoretical, "perfect" ones.

The design of dependently typed interactive theorem provers is a subtle affair. Each new DTITP typically explores new points in the design space, experimenting with novel language features or type theories. This often causes editors, documentation generators, and other tooling to be designed from scratch. This is certainly appealing from a research perspective, but not always the best choice. We recommend that DTITP maintainers consider compromising on their design vision in order to produce more usable systems overall.

Many DTITPs are developed by researchers and lack substantial backing for engineering work and maintenance. This means that when a tool in the ecosystem has reached the "proof of concept" stage, it is not always possible to further develop and maintain it. While this might be enough for initial adoption, it leaves no space for additional improvements and features. Even though DTITPs are different from mainstream programming languages, they *are* still programming languages. This means that solutions to ecosystem requirements already exist, such as the Language Server Protocol (LSP), which can be used to get editor support without having to build that editor from scratch. In the cases where having the "perfect" solution is not feasible, we recommend using an existing "good" solution, since developers from other communities can more easily contribute to its improvement and maintenance.

Similarly, we recommend that the various DTITP communities learn from each other. Multiple participants complained about something being easy in one system, but close to impossible in another.

> [P2] Yeah, I mean, I started out with [Rocq] … and there's a lot of automation of tactics that can be used to automate things. And I used to appreciate that about [Rocq] and I used to miss it a little bit when I was programming in Agda.

We have already seen success from such cross-fertilisation between different DTITPs. For example, the dependently typed pattern matching in Agda has inspired the Equations framework in Rocq [49, 50]. The work in Idris [9] on elaborator reflection [12] was ported to Agda, replacing the previous meta-programming features [53]. Venues, such as the *Workshop on Implementation of Type Systems* (WITS), enable developers of different systems to share their experience and best practices. We believe that expanding such collaboration and cross-fertilisation will lead to better systems in general.

R3: Make sure that users can rely on existing components and standard libraries to stay stable.

The purpose of a standard library is to reduce the amount of work programmers have to do by defining common types and functions for them. This is why we were surprised to learn some participants using Agda actively tried to avoid using the standard library, voicing their concerns about backwards compatibility.

> [P2] So, for example, if you want to apply, well isomorphisms, I think there are like three different notions of isomorphisms … in the standard library and you should use a particular one. And that particular one has changed recently … and then all of the functions surrounding it changed.

If using a definition from the standard library means the programmer has to rewrite a portion of their code the next time they update their DTITP, one of two things will happen: the programmer will not update their DTITP, or the programmer will not use the standard library. The way this problem showed up in our interviews is through the standard library, but a similar reasoning holds

for other tooling. Of course, it can happen that a breaking update is necessary, in which case the users should be informed before updating. For libraries specifically, a common approach used by many mainstream programming languages is to mark unstable definitions explicitly, such that users know that the usage of these definitions leads to potential refactorings later. We recommend that library and tool maintainers put emphasis on remaining stable unless absolutely necessary, and maintain clear communication about breaking changes.

R4: Help new users join the community and understand the relevance of your system.

All three DTITPs already have active communities, and our participants were almost all positive about them. As a result, we expect these DTITPs to become better over time through contributions from their community members. We do, however, believe that larger, more diverse communities can help DTITPs improve faster. In order to achieve community growth, we have two focus points: accessibility and education.

Accessibility. The beginner participants all encountered at least one problem they were not able to solve. Some immediately opted to ask an AI assistant, but some attempted to understand the error message or consult the official documentation, which often proved to be a fruitless endeavour. This suggests that there is a deeper issue around the *accessibility* of these systems. The error messages, documentation, and libraries typically target fellow experts, rather than beginners. Similarly, contributing to DTITP tools can be challenging, since the repositories often lack READMEs and contribution guides, and the code is only sparsely commented and has a non-standard or under-documented architecture. Improving the overall accessibility of these resources may help grow and diversify the user base.

Education. Our other recommendation for community growth is for teachers to provide compelling use cases that illustrate the benefits of these systems. Since our beginner participants were students who came across Agda in a course within their Computer Science bachelor study, we expected them to have a good idea of how they could apply DTITPs in potential future careers as software engineers. This proved to not be the case, and many of them struggled to grasp how these systems could be used in practice. We consider helping potential new users understand the relevance and value of DTITPs a crucial step towards expanding the language community.

R5: Provide an *accepted* and *native* way to opt out of totality.

Developing software in a dependently typed language, even without writing any proofs, can take substantially more time and effort than it would in a simply typed programming language. One of the main factors contributing to this overhead is the fact that DTITPs typically require all functions to be *total*[6] to ensure that the underlying logic is sound. Yet, any real world application is full of partial functions, such as parsing input data, array indexing, integer division, or numerous other (effectful) computations that may fail dynamically. In many cases, these *can* be made total easily enough, for example, by explicitly working with the `Maybe` type or by passing preconditions as additional arguments. Each of these solutions, however, comes at a very real development cost: callers of these partial functions must handle exceptions or prove that they cannot occur. Many projects in DTITPs therefore focus on working on the subset of the problem domain where all functions are total, with the rest usually being delegated to another language. As a result, there is very little language support or interest in working with partial functions, even if they are indispensable for practical purposes.

[6] A total function requires every input from the domain to produce an output in the codomain.

Existing DTITPs do provide some support for defining partial functions. For example, Agda and Lean allow developers to explicitly mark certain functions as terminating, overriding the termination checker, but potentially compromising on the soundness of their solution. Functions written in Rocq or Agda can be extracted to other languages (e.g., OCaml, Scheme, or Haskell), where invoking partial or effectful functions is more commonplace. However, none of our participants mentioned using them in their work and one participant even explained their scepticism on the usability of extraction to Haskell:

> [P1] Like, I think there's [a] perceived potential design pattern where you write code in Agda and you verify it, and you extract the sort of executable part. Um, I don't know. Because … you write an untyped unsafe program, and then you prove (after the fact) that it's safe. It's, I think, the wrong way to do it in a dependently typed language. It's like really your invariants and your code should be tightly merged.

What is lacking in DTITPs is an *accepted* and *native* way to opt out of totality. Current opt-outs, such as compiler directives or postulates, are feasible options but, as the quote above suggests, they do not sit well with some users. The risk of scrutiny from colleagues and reviewers leads DTITP users to avoid such features altogether at the cost of spending considerable time ensuring their entire program is total. We stress that the challenge here is not only technical: such an opt-out need be *accepted* by the community to become effective.

We do not want to discredit the potential effectiveness of code extraction. Extraction to another language, however, introduces an additional layer of indirection between the DTITP and the target language. The overhead required to integrate code in two languages is non-trivial. DTITP maintainers wishing to provide seamless code-extraction features need to spend time customising extraction between these two languages and integrating their tooling with existing tooling for the target language. Users need to learn to work in both the dependently typed language and the target language, where context switching can introduce unnecessary additional cognitive load. While this might be a manageable workload for the current users of extraction, we worry that it is too high for widespread adoption outside of the current user base. This is why we argue for *native* support for totality opt-outs.

We suggest looking towards more mainstream languages which have succeeded in resolving similar tensions between rigour and usability. One approach would be to classify programs into two categories: a verified core and an unverified outer shell, similar to how Haskell encourages writing a pure core and encapsulating all effects in the `IO` monad. Alternatively, partial functions may explicitly be given an exception — much in the way Rust allows code to be marked as unsafe[7]. In both of these cases, the approaches are integrated into the language. The community has accepted guidelines on when and where these "escape hatches" are permissible, and they are presented as features of the respective language in the official documentation.

5 THREATS TO VALIDITY & STUDY LIMITATIONS

Researcher Bias. We have applied the Socio-Technical Grounded Theory methodology to develop theories grounded in qualitative data, and used techniques which inherently integrate our culture, identity, and experiences into the analysis. We acknowledge the constructivist nature of our research paradigm, and ensure the soundness of our work by following well-established guidelines. Based on Hoda's guide to STGT [28, p. 260], our research team regularly reviewed the codebook used during the analysis and discussed the emerging concepts and categories during weekly meetings. Furthermore, as recommended by Hoda [28, pp. 338–339], we provide a detailed overview of how we applied STGT in Section 2, including a description of how participants were recruited, an

[7] Similar to how Agda allows functions to be explicitly marked as terminating when the termination checker fails to prove it.

explanation of how data was collected and analysed, and a demonstration of our coding process with specific examples. We provide our entire codebook as well as the interview transcripts in our public data repository [4] under an MIT Licence. Finally, when presenting the results in Section 3, we "present a clear chain of evidence from [participant] quotations to proposed concepts" [1], as prescribed by the *Empirical Standards for Software Engineering Research*.

Recruitment. We used a combination of convenience sampling and online sign-up to recruit our participants (see Table 1 for the division). While an attempt was made to spread the sign-up survey as wide as possible, its reach was still influenced by our environment. This way of sampling is what lead to the majority of participants being Agda users, which also influenced the direction we took during the STGT iterations. Though we believe to have reached theoretical saturation for the concepts and categories we focused on in our data, a different recruitment strategy might lead researchers towards different avenues of interest.

Interviews. Our dataset consisted entirely of qualitative data in the form of interview transcripts. While this was a suitable way to "find the most [...] possible values of a characteristic in the population", it does mean we are unable to "identify the distribution of characteristic values" in that population [38, p. 1]. While our findings leave us with hypotheses that are grounded in our dataset, future research which applies a combination of qualitative and quantitative techniques could help refine them. For example, it would be interesting to study the prevalence and severity of the positive and negative experiences reported in our work.

Diversity. Only two of our participants were not men, and neither was an expert in using DTITPs. While we do have the impression that this is a representative sample of the *current* DTITP user population, we believe a more diverse participant pool would have provided us with richer insights. We strongly encourage future research to look for a participant pool representative of a *potential* population, if only to determine how best to adjust the object under study to cater to that target.

6 RELATED WORK

Though advanced type systems and interactive theorem provers have been around for some time, user-oriented research has only recently picked up in the field. We discuss the four most relevant studies to our work below:

- Shi et al. [47]'s observation study on how users approach writing proofs in Rocq and Lean;
- Lubin and Chasins [35]'s Grounded Theory on how statically typed functional programmers write code;
- Juhošová et al. [30]'s cross-sectional study on the obstacles new users face when learning to use Agda; and
- their report on developer experience with type-driven development [31].

Similarly to us, Shi et al. [47] conducted think-aloud session with their participants, followed by a short interview in which questions were based on the observations from the think-aloud. Though they focused specifically on proof writing in Rocq and Lean (both tactic-oriented DTITPs), their findings closely reflect ours. They found that their participants often consulted external resources, interacted with the compiler and advanced by incorporating its feedback (H5), considered design aspects beyond getting to a correct proof and refactored their proofs to match a certain style even after it was finished (H4), and struggled with "minutiae" when trying to convince the compiler that they are satisfying the constraints laid down by the types (H7).

Lubin and Chasins [35], whose Grounded Theory study focused on the usage of static type systems, presented hypotheses that are reflected in our own findings. Our observation that refactoring types is a necessary part of writing a DTITP program (H1) is in line with their first hypothesis

stating that functional programmers iterate between editing types and editing expressions. Their second hypothesis, stating that functional programmers run their compilers on code they know will not compile, fits into our theory about conversing with the proof assistant (H5).

Juhošová et al. [31] found that advanced typed systems help users plan and guide their implementation, and allow them to hold an active conversation with the compiler. These findings correspond with our hypotheses on planning (H1) and interactivity (H5). Additionally, their findings about the obstacles that new users face [30] and the improvements that experienced users would like to see [31] support the reasoning behind our recommendations: they find that the coupling between dependent type theory and the language design is a learning obstacle (R1) and that the lack of support for writing real-world applications makes it feel "irrelevant" (R4). A similar study on the usability barriers with liquid types was conducted by Gamboa et al. [19]. While systems with liquid types work differently to DTITPs, the findings in their study are consistent with ours. Specifically, they also identify limited IDE support, lack of reference resources, complex setup, and unhelpful error messages as usability barriers.

Beyond our Recommendations. There is a rich vein of related work, already exploring some of the directions suggested in our recommendations. While giving a complete overview of these areas is outside the scope of this paper, we want to mention some closely related work here.

In the context of functional programming languages, there has been research dating back forty years on providing users with the information they need to identify the problem giving rise to a type error, i.e., on *type error diagnosis* [54]. One approach, taken by Heeren [25] tries to identify the most likely cause of a type error. Alternatively, the approach using *counter-factual* typing suggests the necessary changes to the program to make type correct [11]. Eremondi et al. [17] have suggested how these counter-factual techniques may be applied to the unification algorithm used in dependently typed programming languages specifically.

The "representation problems" we mentioned in R1 are not new: the choice of definition makes some proofs easier than others. The idea of *proof transport* [14] is to enable results from one representation to be carried over to another. Alternatively, providing more control over how definitions are unfolded may help make proofs more robust against definition changes [22].

The design of suitable IDEs for proof assistants is another area of active research. Magnusson and Nordström [36] designed ALF, an influential precursor to Agda, that already supports many of the interactive elements found in today's systems. There have been numerous proposals for new IDEs for Rocq [18, 44, 46], with limited adoption among users. Yet a closely integrated IDE offers new possibilities for programming languages with dependent types, as Christiansen [13] show in their work on *type providers* in Idris.

Usability of Interactive Theorem Provers. There has also been work on the usability of interactive theorem provers which are not necessarily based on dependent types. Aitken et al. [3] evaluated the HOL theorem prover by modelling user activity. Similarly, Beckert and Grebing [7] created and evaluated a model of the user in the proof process. They also evaluated the usability of their own proof assistant [6] using the Cognitive Dimensions framework [24]. They concluded that proof presentation, documentation availability, and overview and management of the proof state are important improvement points for the usability of ITPs. Grebing and Ulbrich [23] evaluate an ITP and give five recommendations for how ITPs should help guide the user. A usability study using focus groups by Beckert et al. [8] found that users believe an intuitive proof process, an understandable proof state, and convenient interaction make for ideal ITPs. Thoma and Iannone [52] showed that different students learning Lean have distinct proving styles. A comparison of two ITP interfaces by Hentschel et al. [26] concluded that debugger-like views are preferrable, at least when the ITP is used to prove properties about an external program. Lapets and Kfoury

[32] designed an interface for an ITP, putting emphasis on library access functionalities within the editor. Another interface design for a different ITP was performed by Mitsch and Platzer [39].

Grounded Theory in Software Engineering. In recent years, Grounded Theory has been applied more in software engineering research [2, 29, 37, 43, 48, 55]. Relevantly, Mugnier et al. [40] used it to study the usage of Dafny [33], a "verification-aware" programming language. Though Dafny uses a different underlying logic than DTITPs, the authors also report inadequate error messages as an obstacle for users (R1) and code style guides as a measure to improve maintainability (H4).

7 CONCLUSION & FUTURE WORK

In this work, we used Socio-Technical Grounded Theory to investigate how people use dependently typed interactive theorem provers. Based on twelve interviews with users of DTITPs, we defined seven hypotheses that form our theory:

(1) DTITP users use types to create a plan for their program and to constrain the solution space. They adjust this plan as necessary throughout the process.
(2) DTITP users aim for a "sweet spot" between intrinsic and extrinsic verification, influenced by the design of the DTITP they are using.
(3) DTITP users search for the "right level" of abstraction in their code.
(4) DTITP users follow a code style geared towards maintainability and community conventions.
(5) DTITP users write code by conversing with the compiler, which can act as an oracle, but also request more information from the user.
(6) More experienced users are able to better respond to the DTITP's requests.
(7) DTITP users want the entire system to behave more like they think.

Based on our findings, we have described five concrete recommendations for anyone who wishes to make DTITP technology accessible to a wider range of software developers:

(1) Keep it simple and design for inexperienced users first.
(2) Prefer existing, "good" solutions over theoretical, "perfect" ones.
(3) Make sure that users can rely on existing components and standard libraries to stay stable.
(4) Help new users join the community and understand the relevance of your system.
(5) Provide an *accepted* and *native* way to opt out of totality.

These recommendations currently specify *what* to focus on, but not necessarily *how* to achieve the desired results. Promising avenues of future research include investigating how to make in-IDE searching of proof libraries better; how error messages can be simplified and made more actionable; and how to better hide the complexities arising from a DTITP's underlying type theory from the user. Additionally, if the goal is to provide everyday developers with access to DTITPs, we believe it is important to investigate how to do so while keeping the solution *practical*, and to consider how DTITPs as a tool fit into the wider context of software engineering. Crucially, we think it is important that any future research in this direction is conducted in a user-oriented rather than theory-driven manner — tools designed for developers should put developer experience first.

We hope that this work helps guide the future development of DTITPs, eventually enabling a wider range of developers to write both *programs and proofs*.

DATA-AVAILABILITY STATEMENT

We provide a MIT-Licenced data repository containing resources for the interview sessions we conducted, anonymised transcripts of those interviews, and our codebook [4].

REFERENCES

[1] 2020. Empirical Standards for Software Engineering Research. https://arxiv.org/abs/2010.03525v2

[2] Steve Adolph, Wendy Hall, and Philippe Kruchten. 2011. Using grounded theory to study the experience of software development. *Empirical Software Engineering* 16, 4 (2011), 487–513. https://doi.org/10.1007/s10664-010-9152-6

[3] J. S. Aitken, P. Gray, T. Melham, and M. Thomas. 1998. Interactive Theorem Proving: An Empirical Study of User Activity. *Journal of Symbolic Computation* 25, 2 (1998), 263–284. https://doi.org/10.1006/jsco.1997.0175

[4] Anonymous Authors. 2026. Dataset for: "A Grounded Theory on the Current State of Dependently Typed Interactive Theorem Proving" (dataset). https://data.4tu.nl/private_datasets/n4dFvuYDozOwa8DfOBrJfB_u8r7Vw0APsNnbdlNOhYI

[5] Brian E. Aydemir, Aaron Bohannon, Matthew Fairbairn, J. Nathan Foster, Benjamin C. Pierce, Peter Sewell, Dimitrios Vytiniotis, Geoffrey Washburn, Stephanie Weirich, and Steve Zdancewic. 2005. Mechanized Metatheory for the Masses: The PoplMark Challenge. In *Theorem Proving in Higher Order Logics*, Joe Hurd and Tom Melham (Eds.). Springer, 50–65. https://doi.org/10.1007/11541868_4

[6] Bernhard Beckert and S. Grebing. 2012. Evaluating the usability of interactive verification systems. *CEUR Workshop* 873 (2012), 3–17.

[7] Bernhard Beckert and Sarah Grebing. 2015. Interactive Theorem Proving - Modelling the User in the Proof Process. https://api.semanticscholar.org/CorpusID:1514923

[8] Bernhard Beckert, Sarah Grebing, and Florian Böhl. 2014. A Usability Evaluation of Interactive Theorem Provers Using Focus Groups. https://doi.org/10.1007/978-3-319-15201-1_1

[9] Edwin Brady. 2017. *Type-Driven Development with Idris*. Manning Publications Co.

[10] Kathy Charmaz. 2014. *Constructing Grounded Theory* (2 ed.). Sage Publications.

[11] Sheng Chen and Martin Erwig. 2014. Counter-factual typing for debugging type errors. In *Symposium on Principles of Programming Languages (POPL '14)*. ACM, 583–594. https://doi.org/10.1145/2535838.2535863

[12] David Christiansen and Edwin Brady. 2016. Elaborator reflection: extending Idris in Idris. In *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming (ICFP 2016)*. Association for Computing Machinery, 284–297. https://doi.org/10.1145/2951913.2951932

[13] David Raymond Christiansen. 2013. Dependent type providers. In *Workshop on Generic Programming (WGP '13)*. ACM, 25–34. https://doi.org/10.1145/2502488.2502495

[14] Cyril Cohen, Enzo Crance, and Assia Mahboubi. 2024. Trocq: Proof Transfer for Free, With or Without Univalence. In *Programming Languages and Systems*, Stephanie Weirich (Ed.). Springer, 239–268. https://doi.org/10.1007/978-3-031-57262-3_10

[15] Juliet Corbin and Anselm Strauss. 2015. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4 ed.). https://uk.sagepub.com/en-gb/eur/basics-of-qualitative-research/book235578

[16] Leonardo de Moura and Ullrich, Sebastian. 2021. The Lean 4 Theorem Prover and Programming Language. In *International Conference on Automated Deduction (CADE 28)*. Springer-Verlag, 625–635. https://doi.org/10.1007/978-3-319-21401-6_26

[17] Joseph Eremondi, Wouter Swierstra, and Jurriaan Hage. 2019. A framework for improving error messages in dependently-typed languages. *Open Computer Science* 9 (2019), 1–32. https://doi.org/10.1515/comp-2019-0001

[18] Alexander Faithfull, Jesper Bengtson, Enrico Tassi, and Carst Tankink. 2018. Coqoon. *International Journal on Software Tools for Technology Transfer* 20, 2 (2018), 125–137. https://doi.org/10.1007/s10009-017-0457-2

[19] Catarina Gamboa, Abigail Reese, Alcides Fonseca, and Jonathan Aldrich. 2025. Usability Barriers for Liquid Types. In *Conference on Programming Language Design and Implementation (PLDI)*. ACM. https://doi.org/10.1145/3729327

[20] Barney G. Glaser. 1978. *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press.

[21] Barney G. Glaser and Anselm L. Strauss. 1967. *The discovery of grounded theory: strategies for qualitative research*. Aldine Publishing, Chicago. OCLC: 253912.

[22] Daniel Gratzer, Jonathan Sterling, Carlo Angiuli, Thierry Coquand, and Lars Birkedal. 2025. Controlling unfolding in type theory. *Mathematical Structures in Computer Science* 35 (2025), e38. https://doi.org/10.1017/S0960129525100327

[23] Sarah Grebing and Mattias Ulbrich. 2020. Usability Recommendations for User Guidance in Deductive Program Verification. In *Deductive Software Verification: Future Perspectives: Reflections on the Occasion of 20 Years of KeY*. Springer International Publishing.

[24] T. R. G. Green and M. Petre. 1996. Usability Analysis of Visual Programming Environments: A 'Cognitive Dimensions' Framework. *Journal of Visual Languages & Computing* 7, 2 (1996), 131–174. https://doi.org/10.1006/jvlc.1996.0009

[25] B. J. Heeren. 2005. *Top quality type error Messages*. Ph. D. Dissertation. Utrecht University. https://dspace.library.uu.nl/handle/1874/7297 Accepted: 2005-09-20T12:43:01Z ISBN: 9789039340059.

[26] Martin Hentschel, Reiner Hähnle, and Richard Bubel. 2016. An empirical evaluation of two user interfaces of an interactive program verifier. In *International Conference on Automated Software Engineering (ASE)*. ACM, 403–413. https://doi.org/10.1145/2970276.2970303

[27] Rashina Hoda. 2022. Socio-Technical Grounded Theory for Software Engineering. *Transactions on Software Engineering* 48, 10 (2022), 3808–3832. https://doi.org/10.1109/TSE.2021.3106280 Conference Name: IEEE Transactions on Software

Engineering.

[28] Rashina Hoda. 2024. *Qualitative Research with Socio-Technical Grounded Theory: A Practical Guide to Qualitative Data Analysis and Theory Development in the Digital World.* Springer Cham. https://doi.org/10.1007/978-3-031-60533-8

[29] Rashina Hoda, James Noble, and Stuart Marshall. 2012. Developing a grounded theory to explain the practices of self-organizing Agile teams. *Empirical Software Engineering* 17, 6 (2012), 609–639. https://doi.org/10.1007/s10664-011-9161-0

[30] Sára Juhošová, Andy Zaidman, and Jesper Cockx. 2025. Pinpointing the Learning Obstacles of an Interactive Theorem Prover. In *International Conference on Program Comprehension (ICPC)*. IEEE, 159–170. https://doi.org/10.1109/ICPC66645.2025.00024

[31] Sára Juhošová, Andy Zaidman, and Jesper Cockx. 2026. The Way of Types: A Report on Developer Experience with Type-Driven Development. In *International Conference on Program Comprehension (ICPC)*. ACM. https://doi.org/10.1145/3794763.3794812

[32] Andrei Lapets and Assaf Kfoury. 2012. A User-friendly Interface for a Lightweight Verification System. *Electronic Notes in Theoretical Computer Science* 285 (2012), 29–41. https://doi.org/10.1016/j.entcs.2012.06.004

[33] K. Rustan M. Leino. 2010. Dafny: an automatic program verifier for functional correctness. In *International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR'10)*. Springer-Verlag, 348–370. https://dl.acm.org/doi/10.5555/1939141.1939161

[34] Xavier Leroy. 2009. Formal verification of a realistic compiler. *Commun. ACM* 52, 7 (2009), 107–115. https://doi.org/10.1145/1538788.1538814

[35] Justin Lubin and Sarah E. Chasins. 2021. How statically-typed functional programmers write code. *Proceedings of the ACM on Programming Languages* 5, OOPSLA (2021), 155:1–155:30. https://doi.org/10.1145/3485532

[36] Lena Magnusson and Bengt Nordström. 1994. The Alf proof editor and its proof engine. In *Types for Proofs and Programs*, Henk Barendregt and Tobias Nipkow (Eds.). Springer, 213–237. https://doi.org/10.1007/3-540-58085-9_78

[37] Zainab Masood, Rashina Hoda, and Kelly Blincoe. 2022. Real World Scrum A Grounded Theory of Variations in Practice. *IEEE Transactions on Software Engineering* 48, 5 (2022), 1579–1591. https://doi.org/10.1109/TSE.2020.3025317 Conference Name: IEEE Transactions on Software Engineering.

[38] Jorge Melegati, Kieran Conboy, and Daniel Graziotin. 2024. Qualitative Surveys in Software Engineering Research: Definition, Critical Review, and Guidelines. *IEEE Transactions on Software Engineering* 50, 12 (2024), 3172–3187. https://doi.org/10.1109/TSE.2024.3474173

[39] Stefan Mitsch and André Platzer. 2017. The KeYmaera X Proof IDE - Concepts on Usability in Hybrid Systems Theorem Proving. *Electronic Proceedings in Theoretical Computer Science* 240 (2017), 67–81. https://doi.org/10.4204/EPTCS.240.5 arXiv:1701.08469 [cs].

[40] Eric Mugnier, Yuanyuan Zhou, Ranjit Jhala, and Michael Coblenz. 2025. On the Impact of Formal Verification on Software Development. *Proceedings of the ACM on Programming Languages* 9, OOPSLA2 (2025), 403:3642–403:3668. https://doi.org/10.1145/3763181

[41] Tobias Nipkow, Markus Wenzel, and Lawrence C. Paulson. 2002. *Isabelle/HOL: a proof assistant for higher-order logic.* Springer-Verlag, Berlin, Heidelberg.

[42] Ulf Norell. 2007. *Towards a practical programming language based on dependent type theory.* PhD Thesis. Chalmers University of Technology, Göteborg, Sweden.

[43] Aastha Pant, Rashina Hoda, Chakkrit Tantithamthavorn, and Burak Turhan. 2024. Ethics in AI through the Practitioner's View: A Grounded Theory Literature Review. https://doi.org/10.48550/arXiv.2206.09514 arXiv:2206.09514 [cs].

[44] Clément Pit-Claudel and Pierre Courtieu. 2016. Company-Coq: Taking Proof General one step closer to a real IDE. In *International Workshop on Coq for PL (CoqPL)*. https://dspace.mit.edu/handle/1721.1/101149.2 Accepted: 2021-09-23T17:56:26Z.

[45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/arXiv.2212.04356 arXiv:2212.04356 [eess].

[46] Kenneth Roe and Scott Smith. 2016. CoqPIE: An IDE Aimed at Improving Proof Development Productivity. In *Interactive Theorem Proving (Lecture Notes in Computer Science)*, Jasmin Christian Blanchette and Stephan Merz (Eds.). Springer International Publishing, Cham, 491–499. https://doi.org/10.1007/978-3-319-43144-4_32

[47] Jessica Shi, Cassia Torczon, Harrison Goldstein, Benjamin C. Pierce, and Andrew Head. 2025. QED in Context: An Observation Study of Proof Assistant Users. *Artifact for QED in Context: An Observation Study of Proof Assistant Users* 9, OOPSLA1 (2025), 92:337–92:363. https://doi.org/10.1145/3720426

[48] Leonardo Sousa, Anderson Oliveira, Willian Oizumi, Simone Barbosa, Alessandro Garcia, Jaejoon Lee, Marcos Kalinowski, Rafael de Mello, Baldoino Fonseca, Roberto Oliveira, Carlos Lucena, and Rodrigo Paes. 2018. Identifying design problems in the source code: a grounded theory. In *International Conference on Software Engineering (ICSE) (ICSE '18)*. ACM, 921–931. https://doi.org/10.1145/3180155.3180239

[49] Matthieu Sozeau. 2010. Equations: A Dependent Pattern-Matching Compiler. In *Interactive Theorem Proving*, Matt Kaufmann and Lawrence C. Paulson (Eds.). Springer, 419–434. https://doi.org/10.1007/978-3-642-14052-5_29

[50] Matthieu Sozeau and Cyprien Mangin. 2019. Equations reloaded: high-level dependently-typed functional programming and proving in Coq. *Equations Reloaded Accompanying Material* 3, ICFP (2019), 86:1–86:29. https://doi.org/10.1145/3341690

[51] The Coq Development Team. 2024. The Coq Proof Assistant. https://zenodo.org/records/14542673

[52] Athina Thoma and Paola Iannone. 2023. Natural Number Game: Students' activity using an interactive theorem prover. In *Congress of the European Society for Research in Mathematics Education (CERME13)*, Vol. TWG14. Alfréd Rényi Institute of Mathematics. https://hal.science/hal-04410448 Backup Publisher: Alfréd Rényi Institute of Mathematics and Eötvös Loránd University of Budapest.

[53] Paul van der Walt and Wouter Swierstra. 2013. Engineering Proof by Reflection in Agda. In *Implementation and Application of Functional Languages*, Ralf Hinze (Ed.). Springer, 157–173. https://doi.org/10.1007/978-3-642-41582-1_10

[54] Mitchell Wand. 1986. Finding the source of type errors. In *Symposium on Principles of programming languages (POPL '86)*. ACM, 38–43. https://doi.org/10.1145/512644.512648

[55] Michael Waterman, James Noble, and George Allan. 2015. How Much Up-Front? A Grounded theory of Agile Architecture. In *International Conference on Software Engineering (ICSE)*, Vol. 1. IEEE, 347–357. https://doi.org/10.1109/ICSE.2015.54