Abstract—Formal methods provide remarkable tools allowing for high levels of confidence in the correctness of developments. Their use is therefore encouraged, when not required, for the development of systems in which safety or security is mandatory. But effectively specifying a secure system or deriving a secure implementation can be tricky. We propose a review of some classical ‘gotchas’ and other possible sources of concerns with the objective to improve the confidence in formal developments, or at least to better assess the actual confidence level.

I. INTRODUCTION

Formal methods applied to the development of systems or software are very efficient tools that allow for high levels of assurance in the validity of the results. By defining languages with clear semantics and by making explicit how to reason on these languages, they provide a mathematical framework in which it is possible to ensure the correctness of implementations. Formal guarantees are often unreachable by more classical approaches; for example they are exhaustive whereas tests cover only a part of the possible executions.

For these reasons, the use of formal methods is encouraged, when not required, by standards for the development of systems in which safety is mandatory, e.g. IEC 61508 [1]. The situation is similar for the development of secure systems: for the highest levels of assurance the Common Criteria (CC, [2]) require the use of formal methods to improve confidence in the development, as well as to ease the independent evaluation process. Indeed, the verification that the delivered product complies with its specification is expected to rely, at least to some extent, on a mechanically checked proof of correctness.

One should not however confuse safety with security. They are overlapping but none includes the other. Safety mostly aims at limiting consequences of random events (dealing with probabilities) and security at managing malicious actions (dealing with the difficulty of an attack). In this paper, we discuss a few concerns more specifically related to the formal development of secure systems. These concerns are illustrated through simple examples (sometimes involving a malicious developer) in Coq [3] or in B [4] but most of them are relevant for other deductive formal methods such as FoCal [5], PVS [6], Isabelle/HOL [7], etc.

This work is supported in part by the Agence Nationale de la Recherche under grant ANR-06-SETI-016 for the SSURF Project.

II. FORMAL METHODS

Standard development processes identify several phases such as specification, design, implementation and verification operations. Different languages can be used for different phases; beyond programming languages it is frequent to use natural language, automata, graphical languages, UML, etc. The problem of the correctness of a development can then be seen as a problem of traceability between the various descriptions of the system produced at different phases.

Formal methods considered in this paper also allow for multiple descriptions of a system; they differ from standard approaches by enforcing the use of languages with explicit and clear semantics, and by providing a logical framework to reason on them. Ensuring the correctness then becomes a mathematical analysis of the traceability (or consistency).

A. About formal specification

At least two descriptions of a system are generally considered in formal methods, a formal specification and an implementation. The specification is often written in a logical language (e.g. based on predicates) and is ideally declarative, abstract, high-level and possibly non-deterministic, describing the what. On the other hand, the implementation is imperative, concrete, low-level, and deterministic, describing the how. To emphasise the difference between declarative and imperative approaches, consider the specification of the integer square root function, \( \sqrt{n^2} \leq n < (\sqrt{n} + 1)^2 \), which is deterministic (for any \( n \) there is at most one acceptable value for \( \sqrt{n} \)) but is not a program: how it is computed is left to the developer.

Simply writing a formal specification is already an improvement compared to standard approaches. Indeed, by using a formal language, ambiguities are resolved. Furthermore, formal methods provide ways to check, at least partially, the consistency of the specification.

B. About refinement

The process of going from a specification to an implementation, while checking the compliance, is called refinement. This concept captures the activity of designing a system; it encompasses a lot of subtle activities, including choosing concrete representations for abstract data or producing operational algorithms matching declarative descriptions. From a logical point of view, a formal specification describes a
family of models (that is, intuitively, implementations) and the refinement process consists of choosing one of those models. Any formal method defines, implicitly or explicitly, a form of refinement. The definition generally ensures that the refinement is transitive (allowing for an arbitrary number of refinement steps in the development process) and monotone (allowing for the decomposition of a problem into several sub-problems that will be refined independently).

Formal methods do not automatically produce refinements but explain how to check that a refinement is valid, that is they ensure that very different objects (a logical description and an operational implementation) are sufficiently 'similar'. To allow for the decomposition of a problem into several sub-problems that will be refined independently.

The notion of refinement is expressed between machines (modules combining a state defined by variables, properties such as an invariant on the state and operations encoded as substitutions to read or alter the state) and captures the essence of program correctness w.r.t. their specification as follows: an implementation refines a specification if the user cannot exhibit a behaviour of the implementation that is not compliant with what is required by the specification. This concept is incorporated into the methodology by the automated generation of proof obligations at each refinement step, and is sustained by mathematical justifications not detailed here.

One of the characteristics of the refinement in B is that it is independent of the internal representation used by the machines, as illustrated by the following example of a system returning the maximum of a set of stored natural values:

\[
\text{MACHINE } M_A \\
\text{VARIABLES } S \\
\text{INVARIANT } S \subseteq \mathbb{N} \\
\text{INITIALISATION } S := \emptyset \\
\text{OPERATIONS} \\
\quad \text{store}(n) \triangleq \text{PRE } n \in \mathbb{N} \text{ THEN } S := S \cup \{n\} \\
\quad m \leftarrow \text{get} \triangleq \text{PRE } S \neq \emptyset \text{ THEN } m := \max(S) \]

\[
\text{END} \\
\text{MACHINE } M_C \text{ REFINES } M_A \\
\text{VARIABLES } s \\
\text{INVARIANT } s = \max(S \cup \{0\}) \\
\text{INITIALISATION } s := 0 \\
\text{OPERATIONS} \\
\quad \text{store}(n) \triangleq \text{IF } s < n \text{ THEN } s := n \\
\quad m \leftarrow \text{get} \triangleq m := s \]

The state of the machines is described in the VARIABLES clause; for the specification \(M_A\) it is a set of natural numbers and for the implementation \(M_C\) a natural number. The INVARIANT clause defines a constraint over the state; for \(M_A\) it indicates that \(S\) is a subset of \(\mathbb{N}\), whereas for \(M_C\) it describes the glue between the states of \(M_A\) and \(M_C\) (intuitively claiming that if both machines are used in parallel then \(s\) is always equal to \(\max(S)\)). The INITIALISATION clause sets the initial state, while the OPERATIONS clause details the operations used to read or alter the state. The two machines differ yet \(M_C\) refines \(M_A\): roughly speaking one cannot exhibit a property of \(M_C\) which contradicts one of \(M_A\).

Note the use of the PRE substitution defining a precondition, that is a condition that the user has to check before calling an operation. This is an offensive approach; an operation (should not but) can be used when this condition is not satisfied, yet in such a case there is no guarantee about the
result (it may even cause a crash). By opposition the **defensive** approach is represented in \( B \) by using **guards** (that is an **IF**) that prevent unauthorised uses. These notions are standard in formal methods and will be discussed further later in this paper.

**B. About Coq**

*C* is a proof assistant based on a type theory. It offers a higher-order logical framework that allows for the construction and verification of proofs, as well as the development and analysis of functional programs in an \( ML \) like language with pattern-matching.

*C* implements the *Calculus of Inductive Constructions* [9] and it is frequent in developments to use inductive definitions. For example, \( N \) is defined in the *Peano* style as follows:

\[
\text{Inductive } N := 0 : N \mid S : N \rightarrow N
\]

This definition means that \( N \) is the smallest set of terms closed under (finite number of) applications of the *constructors* 0 and \( S \). \( N \) is thus made of the terms 0 and \( S^n(0) \) for any finite \( n \); being well-founded, structural induction on \( N \) is possible (the induction principle is automatically derived by *Coq* after the definition of \( N \)). The definition also means that \( N \) contains no other values (**surjectivity**) and that \( \forall (n : N), \ 0 \neq S(n) \) and \( \forall (m n : N), \ S(m) = S(n) \Rightarrow m = n \ (**injectivity**). Contrary to \( B \), there is no enforced development methodology in *Coq*, nor any explicit refinement process. The user can choose between several styles of specification and implementation, and has to decide on its own about the properties to be checked. For example the **weak specification** style consists of defining functions as programs in the internal \( ML \)-like language and later checking properties of these functions, as illustrated here by the division by 2:

\[
\text{Fixpoint div2(x : N) : N :=}
\]

\[
\text{match x with S(S(x'))} \rightarrow S(div2(x')) \ | \ _ \ \rightarrow 0 \ \text{end.}
\]

\[
\text{Theorem div2_def :}
\]

\[
\forall all (x : N), \ n = 2 \times \text{div2}(n) \lor n = 2 \times \text{div2}(n) + 1.
\]

\[
\text{Proof.}
\]

\[
\text{Qed.}
\]

div2 is a recursive program (using \( \text{div2}(x+2) = \text{div2}(x)+1 \)) and \( \text{div2_def} \) a property claimed about it; the proof, not detailed here, ensures that \( \text{div2} \) indeed satisfies \( \text{div2_def} \).

**IV. Specifying Secure Systems**

We now begin our discussion about developing secure systems using formal methods by considering more specifically formal specifications of secure systems.

To start with trivial considerations, we first have to note that formal methods offer tools to express specifications but that there is no way to force a developer to describe the properties required of the system under development. Clearly, using even the most efficient formal method without adopting the ‘formal spirit’ is meaningless, as there is no benefit compared to standard approaches if the formal specification is empty. Note also that a formal development is a development, and so can also benefit from standard practices such as naming conventions, modularity, documentation, etc. In the case of formal methods, in fact, the very process of deriving a formal specification from the book of specifications should be documented, justifying the formalisation choices and identifying, if any, aspects of the system left out (as it is generally not reasonable or even feasible to aim at a full formalisation of a complete system).

Assuming a developer that has adopted the formal spirit, there are further points to care about in order to develop an ‘adequate’ formal specification for a secure system, that is a specification not only expressing the required properties, but also ensuring that those properties are enforced at all stages of the development as well as in any (reasonable) scenario of usage of the implementation.

Some of the concerns that will be discussed below are applicable for safety or any high assurance system; for others a malicious developer will be assumed (a threat generally irrelevant for safety but applicable in security). The ultimate objective of such a malicious developer is to exploit any weakness of a specification, in order to trap a system while delivering a mechanically checked proof of compliance. One could consider that such traps would be detected through code review or testing. Yet, beyond the fact that formal methods are expected to reduce the need for such activities, we warn the reader that our illustrations are voluntarily simplistic, and that real life examples of *Trojan Horse* are difficult to detect.

**A. About invalid specifications**

As pointed out in Par. II-C, inconsistent specifications are disastrous. Indeed, whereas inconsistency cannot be automatically detected, it also permits to discharge any proof obligation expressed – that is an inconsistent specification can in practice make the developer life more comfortable. An inconsistent specification is therefore dangerous for safety developments if a distracted developer fails to notice that its proofs are a little too easy to produce, and more so for security developments as a malicious developer identifying such a flaw would be able to prove whatever he wants.

Of course, an inconsistent specification is not implementable. It is therefore possible to check the consistency by providing an implementation – any one will do the trick, so even a dummy implementation is sufficient. Yet there are inconsistencies not detectable in the security situations in which a formal specification is mandatory while a formal implementation is not. This is the case for the *CC*, at some assurance levels, that just require a formal specification of the *Security Policy*. An undetected inconsistent specification is therefore a possibility.

In \( B \) the consistency of a specification is partially checked through proof obligations to be discharged by the developer. Yet the obligations related to the existence of values satisfying the expressed constraints for parameters, variables and constants are deferred. Both following specifications are inconsistent, yet all *explicit* proofs obligations can be discharged
(that is, most $B$ tools will report a ‘100% proven’ status):

\[
\begin{align*}
\text{MACHINE absurd_var} & \quad \text{MACHINE absurd_cst} \\
\text{VARIABLES } v & \quad \text{CONSTANTS } f \\
\text{INVARIANT } & \quad \text{PROPERTIES} \\
& v \in \mathbb{N} \land f \in \mathbb{N} \land \\
& v = 0 \land v = 1 \quad \forall x, y, x < y \Rightarrow f(x) > f(y) \\
\text{ASSERTION } 0 = 1 & \quad \text{ASSERTION } 0 = 1
\end{align*}
\]

Of course, delaying such proof obligations is justified, as implementing the specification will force the developer to exhibit a witness for $v$ that meets the specification (a constructive proof that the specification is satisfiable). Therefore, $B$ ensures that any inconsistency is detected, at the latest, at the implementation stage. But we would like to remind the reader that a formally derived implementation is not always required. In such a case, one should consider additional manual verifications to check the existence of valid values for parameters, constants and variables.

Inconsistencies can be rather easy to introduce, accidentally or not, by contradicting implicit hypotheses associated to the used formal method. In $B$ for example there is a clause $\text{SETS}$ that allows for the declaration of abstract sets used in a machine; one can easily forget that such a set is always in $B$ finite and non-empty. If the developer contradicts one of these implicit hypotheses the specification becomes inconsistent without any warning by the tool; in fact the automated prover will very efficiently detect the contradiction as a lemma usable to discharge any proof obligation. Contradiction of implicit principles of the underlying logic can also be illustrated in Coq with two very simple examples. The first one is a naive tentative of specifying $\mathbb{Z}$ using $\mathbb{N}$:

\[
\begin{align*}
\text{Inductive } & \quad \text{Hypothesis} \\
\mathbb{Z} : \text{Set} := plus : \mathbb{N} \rightarrow \mathbb{Z} & \quad \text{zero unsigned : } \mathbb{N} \rightarrow \mathbb{Z}. \\
\text{minus} : \mathbb{N} \rightarrow \mathbb{Z}.
\end{align*}
\]

Unfortunately, as pointed out in Par. III-B, the definition of $\mathbb{Z}$ is not a specification but an implementation ($\mathbb{Z}$ is the set of all terms of the form $\text{plus}(n)$ or $\text{minus}(n)$). $\text{zero unsigned}$ introduces an inconsistency because it contradicts the injectivity principle for the constructors: for any natural values $n$ and $m$ it is possible to prove in Coq that $\text{plus}(n) \neq \text{minus}(m)$.

The second example is related to the unexpected consequences of using possibly empty types. This is illustrated by the following (missed) attempt to define bi-colored lists of natural values, that is lists with each element marked red or blue:

\[
\begin{align*}
\text{Inductive } & \quad \text{Introductory} \\
\text{blst : Set := red : blst } \rightarrow \mathbb{N} & \rightarrow \text{blst} \\
& | \quad \text{blue : blst } \rightarrow \mathbb{N} \rightarrow \text{blst}.
\end{align*}
\]

In the absence of an atomic constructor for the empty list, $\text{blst}$ which is the smallest set of terms stable by application of the constructors is indeed empty. Therefore, assuming the existence of such a list is inconsistent, and any theorem of the form $\forall (b : \text{blst})$, $P$ is provable – hardly a problem from the developer’s point of view, as he generally tries to prove only those properties he expects. It would be prudent for any type $T$ introduced in Coq, to ensure that it is not empty e.g. by proving a theorem of the form $\exists (t : T)$, True.

One could also investigate the satisfiability of the preconditions or guards, as defined in Par. III-A, associated to functions or operations. Indeed, while unsatisfiable preconditions are not inconsistent, they often represent a form of deadlock, as they mean that it is never possible to use an operation. They may however be difficult to detect – there is a famous example of the database of individuals developed in [4], in which it is impossible to insert new entries, as pointed out in [10], due to the fact that any new individual introduced in the database should have a father and a mother, while the initial state is an empty database. To avoid such difficulties the use of adequate tools (animation of models, model-checking, automatic tests generator, cf. [11]–[15]) can be of considerable help.

We would also like to draw the attention of the readers to other types of problematic specifications. For example in some cases it may happen that a specification mixes predicates of the form $P \Rightarrow Q$ and $P \Rightarrow \neg Q$. Such a specification is consistent but only as long as $P$ is false; to the least this type of specification should be considered inappropriate. This is one of the cases for which specification engineering tools would be considered useful. Such tools associate for example to a specification $\forall x, P \Rightarrow Q$ an additional proof obligation $\exists x, P$; indeed the specification can be vacuously true if $P$ is always false, but it is unlikely that such a specification convey the intended meaning [16].

B. About (mis)understandings

Consequences of invalid specifications have been identified and justify establishing procedures to check consistency. We now discuss the problem of insufficient specifications, which is more tricky to detect as it generally refers to a difference between a specification and its intended meaning.

Our very first concern is related to the understanding of the chosen formal method. It is not reasonable to expect all ‘users’ of formal methods to be expert. One may consider for example a situation in which a customer convinced by the interest of formal methods may however not have any in-depth knowledge about any of them. In fact, we would also argue that should formal methods be more widely used – definitely something we expect for the future – they should be accessible to people having received a dedicated training but which are not expert (this is one of the main objectives of the FoCal project [5], [17]–[19]). The minimum, however, is to ensure that any user has a basic understanding of some of the underlying principles to avoid misinterpretation.

For example, consider the concept of refinement as introduced in Par. II-B. The essence of this concept is to allow to check that specifications and implementations are `similar’. This similarity should not be too strong, as a refinement relation reduced to intensional equality of programs (that is, the same code) would be useless. It is for example standard to consider that computations and transient states are irrelevant. In Coq this is translated by the fact that the equality is modulo $\beta$-reduction (in other words, $\text{square}(3) = 9$ because computing $\text{square}(3)$ yields 9). Our concern is illustrated in $B$ by the
following specification of an airlock system:

MACHINE Sas
VARIABLES door₁, door₂
INVARIANT \(\text{door₁, door₂} \in \{\text{open, locked}\} \land \neg \text{(door₁ = open} \land \text{door₂ = open)}\)

OPERATIONS
\(\text{open₁} \triangleq \text{IF door₂ = locked THEN door₁ := open; ELSE door₁ := locked}\)
\(\text{close₁} \triangleq \text{door₁ := locked}\)
\(\text{open₂} \triangleq \text{IF door₁ = locked THEN door₂ := open; ELSE door₂ := locked}\)

If the underlying principles of the B are not understood, one can easily consider that the INVARIANT clause in a proven B machine is ‘always true’. Therefore, any compliant implementation of this specification would be considered safe. Of course, this is not the case, as we may for example refine the operation \(\text{open₁}\) as follows:

\(\text{open₁} \triangleq \text{IF door₂ = locked THEN door₁ := open; ELSE door₁ := locked}\)

where \text{wait} is a passive but slow operation and \text{attack} any condition the malicious developer can imagine to obfuscate the dangerous behaviour during tests.

If stronger forms of invariant are required, e.g. to take into account interruptions, specific modelisation choices or dedicated techniques are to be used (cf. [20]).

C. About partial specifications

Another aspect of a formal specification of a secure system to check is totality: is the behaviour of the system specified in any possible circumstance? It is frequent in formal methods to define partial specifications – either to represent a form of contract (a condition to be realised before having the right to use the system) or a form of freedom left to the developer (because the system is not planned to be used in such conditions or because the result is irrelevant). If the first interpretation can be considered during formal developments, the second one becomes the only relevant one once leaving the abstract world of formal methods to tackle with implemented systems. And the extent of the freedom given to the developer is easily underestimated, as illustrated in the following examples.

We start by two specifications of the \text{head} function (returning the first element of a list of natural values) in Coq, in the strong specification style\(^1\):

\[
\text{head₁}((l: \text{list } \mathbb{N})(p: l \neq [])) : \{ x: \mathbb{N} \mid \exists l' : \text{list } \mathbb{N}, l = x :: l' \}.
\]

\[
\text{head₂}((l: \text{list } \mathbb{N})) : \{ x: \mathbb{N} \mid l \neq [] \rightarrow \exists l' : \text{list } \mathbb{N}, l = x :: l' \}.
\]

Both specifications ensure that the function, called upon a non empty list, will return the head element. Yet the first specification is associated to a precondition, the parameter \(p\) being a proof that the list parameter \(l\) is not empty – making it impossible to call \text{head₁} over an empty list as it would not be possible to build such a proof. The second specification is on the contrary partial, allowing to use \text{head₂} with an empty list but not constraining the result in such a case (except for being a natural value).

\(^1\)In which the return value of a function is described as satisfying a property.

The point is that these two specifications are not so different: all the logical parts of a Coq development are eliminated at extraction (the process that extract proved programs). This is not specific to Coq: by nature, logical contents in a formal development are not computable and have therefore to be discarded in some way before being able to produce a program. And it is easy to implement both specifications in a way that produces the same following OCaml code, where \text{secret} is any value the malicious developer would care to export:

\[
\text{let head = function } [] \rightarrow \text{ secret } | h ::= \_ \rightarrow h
\]

We illustrate the same concern in B by the specification of a file system manager. We define the sets USR (users), FIL (files), CNT (contents) and RGT (access rights). \text{Cnt} associates for any file a content, \text{Rgt} associates for a user and a file the rights, and \text{cnt} gives the number of existing files. Various operations to create, delete or access the files are assumed to be specified but are not detailed here, except for \text{read}:

MACHINE filesystem
SETS USR; FIL; CNT; RGT = \{ r, w \}
CONSTANTS cnul
PROPERTIES cnul \in CNT
VARIABLES Fil, Cnt, Rgt, cnt
INVARIANT Fil \subseteq FIL \land
\text{Cnt} \in \text{Fil} \land \text{Rgt} \subseteq (\text{USR} \times \text{Fil}) \times \text{RGT} \land
\text{cnt} = \text{card}(\text{Fil})
INITIALISATION Fil := \emptyset \land \text{Cnt} := \emptyset \land \text{Rgt} := \emptyset \land \text{cnt} := 0
OPERATIONS
\dots
\text{out} \leftarrow \text{read}(f, u) \triangleq
\text{PRE} f \in \text{Fil} \land u \in \text{USR} \land
\text{IF } ((u \rightarrow f) \rightarrow r) \in \text{Rgt} \text{ THEN out} ::= \text{Cnt}(f) \text{ ELSE out} ::= \text{cnul}
\dots
\]

\text{read} is specified as returning the content of a file \(f\), provided that the user \(u\) has the right to read it. Yet it is only partially specified, as we do not describe what happens when the file does not exist. Any call of \text{read} implemented in B would be associated to a proof obligation to ensure that the precondition is met, but this constraint goes as far as goes the use of the B. So let’s assume the following malicious refinement of \text{read} is called over a non existing file:

\[
\text{out} \leftarrow \text{read}(f, u) \triangleq
\text{IF } f \in \text{Fil} \text{ THEN}
\text{IF } ((u \rightarrow f) \rightarrow r) \in \text{Rgt} \text{ THEN out} ::= \text{Cnt}(f)
\text{ ELSE out} ::= \text{cnul}
\text{ELSE Fil} := \text{Fil} \cup \{ f_S \} \land
\text{Cnt} ::= \text{Cnt} \cup \{ f_S \rightarrow S \} \land
\text{Rgt} ::= \text{Rgt} \cup \{ (\text{eni} \rightarrow f_S) \rightarrow r \}
\]

Whereas the specification of \text{read} was apparently passive (not modifying the state), this refinement creates a file \(f_S\) storing a (confidential) value \(S\), file only accessible by an arbitrary user \text{eni} invented by the developer. Furthermore the invariant is broken as \(f_S\) is created yet not accounted for in \(\text{cnt}\), that is \(f_S\) is virtually invisible for the system. Note also that defining the returned value when the file does not exist is not even
required by $B$; a malicious developer may however prefer to return $\text{cnul}$ for a better obfuscation of its code.

Clearly, a partial specification cannot enforce security, and one should favor a total (and defensive) specification. In $B$ this would translate into using a IF instead of a PRE. When the condition associated to an IF substitution is not satisfied, the ELSE branch is executed – if it is absent it is equivalent to a skip substitution, that is it enforces to do nothing. On the contrary when the condition associated to a PRE substitution is not satisfied, there is absolutely no guarantee about the result. Note that the defensive approach (with redundant checks) is an implementation of the defence in depth concept.

D. About elusive properties

For our next point, we would like to emphasise that some concepts often encountered in security can be difficult to express in a formal specification. Confidentiality is a good example: while a formal specification may appear to implicitly provide confidentiality, one should be extremely careful about its exact meaning, as illustrated by the following example of a login manager in $B$.

The system state is defined by $\text{Acc} \subseteq \text{USR}$ the accounts, $\text{log}$ to identify the currently logged account (nouser encoding no opened session), and $\text{Pwd}$ to associate to any account a password. This last piece of information is confidential and should not be disclosed. Operations (not detailed in this paper) allows to log, exit, create or destroy an account, with only the log operation specified as depending upon $\text{Pwd}$ to represent the confidentiality of this data. The operation $\text{accounts}$, detailed here, returns the existing accounts:

\[
\text{MACHINE login} \\
\text{SETS USR; PWD} \\
\text{CONSTANTS root, nouser} \\
\text{PROPERTIES root} \in \text{USR} \land \text{nouser} \notin \text{USR} \setminus \{\text{root}\} \\
\text{VARIABLES Acc, log, Pwd} \\
\text{INVARIANT} \\
\text{ Acc} \subseteq \text{USR} \land \text{root} \in \text{Acc} \land \text{nouser} \notin \text{Acc} \land \text{log} \in \text{Acc} \cup \{\text{nouser}\} \land \text{Pwd} \in \text{Acc} \rightarrow \text{PWD} \\
\text{INITIALISATION} \\
\text{ Acc} := \{\text{root}\} \land \text{log} := \text{nouser} \land \text{Pwd} := \{\text{root}\} \rightarrow \text{PWD} \\
\text{OPERATIONS} \\
\text{ out} \leftarrow \text{accounts} \triangleq \\
\text{ IF } \text{log} \in \text{Acc} \text{ THEN} \\
\text{ ANY } s \text{ WHERE } s \in \text{seq(USR)} \land \text{ran}(s) = \text{Acc} \land \text{size}(s) = \text{card}(\text{Acc}) \\
\text{ THEN } \text{out} := s \\
\text{ ELSE out} := \emptyset \\
\text{ }
\text{ }
\text{Input and output values being not refinable in } B \text{ (cf. Par. III-A), the type of the return value of } \text{accounts} \text{ has to be finalised in the specification. In our example, we have chosen to implement the set } \text{Acc} \text{ returned by } \text{accounts} \text{ as a list (or sequence in the } B \text{ terminology) } s \text{ of values of } \text{USR}; \text{ran}(s) = \text{Acc} \text{ ensures that the same values appear in } \text{Acc} \text{ and } s, \text{size}(s) = \text{card}(\text{Acc}) \text{ that the length of the list } s \text{ is equal to the cardinal of } \text{Acc}. \text{The proposed malicious refinement of } \text{accounts} \text{ is the following one:}
\]

\[
\text{ out} \leftarrow \text{accounts} \triangleq \\
\text{ IF } \text{log} \in \text{Acc} \text{ THEN} \\
\text{ ANY } s \text{ WHERE } s \in \text{seq(USR)} \land \text{ran}(s) = \text{Acc} \land \text{size}(s) = \text{card}(\text{Acc}) \\
\text{ THEN } \text{IF } \text{Pwd}(\text{root}) < \text{guess} \text{ THEN wait(10) ELSE wait(20)} \text{ ELSE out} := \emptyset
\]

where $\text{guess}$ is a new variable controlled by the malicious developer. Combining calls to $\text{accounts}$ and changes of $\text{guess}$, one can quickly derive $\text{Pwd}(\text{root})$ through the artificial dependency introduced in the returned value.

This example illustrates a covert channel exploit [21], as discussed in [22]. Even if the implementation stores $\text{Pwd}$ in a private memory location protected by a trusted operating system – a rather optimistic assumption – its confidentiality cannot be guaranteed without a form of control over dependencies (e.g. considering data-flow).

It is of course possible to impose a complete (or monomorphic) specification [23] – a deterministic specification, enforcing the extensional behaviour of any implementation. A complete specification would not let any freedom to the developer and thus would ensure that there is no covert channel to be exploited. In our example, a complete specification would for example require $s$ to be sorted in ascending order. This is however an impractical technical solution, an indirect mean to ensure confidentiality. Furthermore completeness is not expressible in the $B$ specification language (or in most languages considered in this paper), is generally undecidable and is not stable by refinement of the representation of the data – e.g. refining a set by an ordered structure.

It is also possible to better control dependencies in $B$ by specifying operations using constant functions. The following modified specification claims that the operation $\text{accounts}$ behaves like a function depending only upon the set $\text{Acc}$ and returning a list of values of $\text{USR}$:

\[
\text{CONSTANTS . . . . fct} \\
\text{PROPERTIES . . . . } fct \in \mathbb{P}(\text{USR}) \rightarrow \text{seq(USR)} \\
\text{OPERATIONS} \\
\text{ out} \leftarrow \text{accounts} \triangleq \\
\text{ IF } \text{log} \in \text{Acc} \text{ THEN out} := fct(\text{Acc}) \text{ ELSE out} := \emptyset
\]

This approach is not yet fully satisfactory as only the dependencies for the result are described (the extensional point of view). It is therefore still possible to affect the behaviour of $\text{accounts}$, as in this valid refinement:

\[
\text{ out} \leftarrow \text{accounts} \triangleq \\
\text{ out} := \text{encode}(<\text{Pwd}(\text{root}))>; \\
\text{ IF } \text{Pwd}(\text{root}) < \text{guess} \text{ THEN wait(10) ELSE wait(20)}; \\
\text{ IF } \text{log} \in \text{Acc} \text{ THEN out} := fct(\text{Acc}) \text{ ELSE out} := \emptyset
\]

In this refinement the malicious developer implements both a timed channel as well as a possibly observable transient state of the output.

This illustration is just intended to show why, in some cases, expressing confidentiality can be difficult. For such properties,
complementary approaches should be considered, based e.g. on dependency calculus or non-interference [24], [25], and associated to standard code analysis. Note that confidentiality is often formally addressed through access control enforced by a form of monitor, that is according to the Orange Book a tamperproof, unavoidable, and ‘simple enough to be trusted’ mechanism filtering accesses (cf. recent discussions in [26]–[28]). Such a monitor can itself implement this type of covert channel attacks if it is poorly specified. Note also that the confidence in a system implementing a monitor relies on the confidence in the information used by this monitor, such as the source of an access request (that would require a form of authentication) as well as the level of protection required by the accessed object (a meta-information whose origin is generally unclear, but for which effective implementations such as security labels protected in integrity have been proposed).

We mention authentication and integrity to point out another source of rather elusive properties, that is the characterisation of cryptographic functions. For example, a (cryptographic) hash function \( H \) has the following three properties:

- given \( h \) it is not possible to find \( x \) s.t. \( H(x) = h \);
- given \( x \) it is not possible to find \( y \neq x \) s.t. \( H(x) = H(y) \);
- it is not possible to find \( x \neq y \) s.t. \( H(x) = H(y) \).

The first property, for example, guarantees the security of the Unix login scheme; being able to specify a hash function (without giving any details on its implementation) by formally describing these properties has therefore some interest to certify such a scheme. Yet these properties appear to be rather difficult to express formally. A naive translation of the last property would just say that \( H \) is injective, which is false (as \( H \) projects an infinite set in a finite set of binary words of fixed length) and would lead to an inconsistent specification. Formally expressing such properties is possible, but generally less straightforward than one may expect.

**E. About the refinement paradox**

Most of the examples detailed in Pars. IV-C and IV-D are illustrations of what is often referred to as the refinement paradox: some properties are preserved by refinement (safety ones generally are), other are not (security ones).

Back to the discussion of Par. IV-D, the most simple example of ‘devious’ refinement that we can exhibit in \( B \) is the following one:

**MACHINE Boolean**

**OPERATIONS**

\[
\text{out} \leftarrow \text{go} \triangleq \text{IF } \text{dump} \mod 2 = 0 \text{ THEN } \text{out} := \text{true} \text{ ELSE } \text{out} := \text{false};
\]

\[
\text{dump} := \text{dump}/2
\]

One should not believe that the refinement paradox is specific to those methods which are providing an explicit form of refinement, such as \( B \) or \( Z \) for example. Our devious refinements include implicitly a non functional refinement of the representation of data: we accept several implementations as representing a single abstract value of the specification. This intuitively describes why some variables are hidden at the specification level. From this intuition, we suggest the following counterpart in Coq of the refinement paradox. Let’s consider the example of the specification of booleans as an Abstract Data Type, with the equality and a boolean function:

**Module Type Boolean Function.**

**Parameter B : Set.**

**Parameters** \( \top \perp : B. \)

**Hypothesis refl : \forall (b : B). b \equiv b.\)**

**Hypothesis sym : \forall (b1 b2 : B). b1 \equiv b2 \equiv b2 \equiv b1.\)**

**Hypothesis tran : \forall (b1 b2 b3 : B). b1 \equiv b2 \equiv b2 \equiv b3 \equiv b1 \equiv b3.\)**

**Hypothesis inj : \neg \top \equiv \bot.\)**

**Hypothesis surj : \forall (b : B). b \equiv \top \lor b \equiv \bot.\)**

**Parameter fnc : \forall B : B.\)**

End Boolean Function.

The straightforward refinement of this specification is of course to implement \( B \) as \( B \), the Coq type of booleans, and to implement \( \text{fnc} \) as one of the four possible boolean functions (\( \text{true}, \text{false}, \text{identity} \) or \( \text{not} \)). But a devious implementation gives much more freedom; we can for example choose to implement \( B \) as \( \mathbb{N} \), even values representing \( \bot \) and odd values representing \( \top \):

**Module Covert Channel : Boolean Function.**

**Definition B := \mathbb{N}.\)**

**Definition \( \bot := 0. \)**

**Definition \( \top := 1. \)**

**Definition \( \equiv (b1 b2 : B) := (b1 + b2 \mod 2 = 0). \)**

\[
\text{Def}_{\text{fnc}}(b : B) := \text{match } ((b/2) \mod 4) \text{ with }
\]

\[
| 0 \Rightarrow \bot \\
| 1 \Rightarrow \top \\
| 2 \Rightarrow b \\
| \bot \Rightarrow b+1
\]

end.

End Covert Channel.

This implementation introduces a new dimension in the representation of the data, which is hidden at specification level and can be used by a malicious developer to store information and modify results: \text{fnc} now emulates any of the boolean functions.

Note that the term of refinement paradox may be considered an overstatement, provided the presentation of refinement in Par. II-B. Clearly the very concept of refinement is extensional,
whereas on the contrary confidentiality can be considered as intentional: rather than describing what a result should be, it aims at constraining how a result is produced (in this case, without depending upon the confidential value). Similarly, if refinement is intended to preserve properties described in a specification, it does not aim at preserving properties of the specification itself, or any other form of meta-properties; so the fact that for example completeness is not preserved should not be a surprise.

V. Building on Sand?

In Par. IV-A, we have shown possible consequences of inconsistent specifications. Obviously similar or worse consequences can result from other sources of inconsistencies, such as a bug in the tool implementing the formal method, or a mistake in the theory of the formal method itself. For a malicious developer, a paradox (a flaw in the logic that can be used to prove at the same time both $P$ and $\neg P$) discovered in a theory or in a tool can be used to prove any property about any development, that is to implement any unpleasant behaviour while getting a certification.

When trying to assess the level of confidence one may have in the result of a formal development, the question of the validity of the tool and of the theory should therefore be addressed.

A. About the logic

In [29] a deep embedding (cf. [30], [31]) of the B logic in Coq is described, that is intuitively a form of B virtual machine developed in Coq with the objective to check the validity of the B logic. While this deep embedding has not identified any paradox, it has shown that the following ‘theorems’ from [4] are in fact not provable using the defined logic:

\[
E_1 \implies F_3 = E_2 \implies F_2 \implies E_1 = E_2 \\
E_1 \implies F_3 = E_2 \implies F_2 \implies F_1 = F_2 \\
S_1 \subseteq S_2 \land T_1 \subseteq T_2 \implies S_1 \times T_1 \subseteq S_2 \times T_2
\]

These results are not provable because of the definition of the B inference rules, which are not sufficiently precise regarding the formal definition of what is a cartesian product. To our knowledge, the fact that these results were not valid in B was not known by the B community. Being apparently trivial, they were never checked and have been integrated for example in provers for the B logic. That means, at a fundamental level, that these results were in fact taken as additional axioms, without people knowing it – an approach that could have created a paradox in the logic.

Further investigations have emphasised another form of subtile glitch that may appear in the theory of a formal method. As pointed out in Par. II-A, formal methods allow for multiple descriptions of a system as well as the verification of the similarity of these descriptions. This is sometimes obtained by defining several semantics for a single construct.

In B, substitutions of the GSL (used to write operations) are defined as predicate transformers, that is a logical semantic. On the other hand the substitutions of the

\[
B_0
\]

sub-language are used for implementation and also have an operational semantic. This is the case of the WHILE $P$ DO $S$ INVARIANT $I$ / VARIANT $V$ substitution, illustrated in [4] by the extraction of the minimum of a non-empty set of natural values:

\[
\begin{align*}
\text{WHILE } x \notin S \\
\text{DO } x := x + 1 \\
x := 0; \quad \text{INVARIANT } x \in [0, \min(S)] \\
\text{VARIANT } \min(S) \rightarrow x \\
\text{END}
\end{align*}
\]

Using the definition of the WHILE substitution as a predicate transformer, one can indeed show that this substitution realises (that is, transforms into a tautology) the predicate $x = \min(S)$. In other words the substitution is proven to extract the minimum in any case of use (provided $S \neq \emptyset$).

By denoting $\llbracket P \rrbracket$ the translation producing a $C$ program from a $B_0$ substitution, the operational semantic is defined by:

\[
\llbracket \text{WHILE } P \text{ DO } S \text{ INVARIANT } I \text{ / VARIANT } V \rrbracket = \text{while } \llbracket P \rrbracket \{[S]\}
\]

The interesting point is that this semantic forgets $I$ (the loop invariant) and $V$ (the loop variant) that are pure logical contents, important for the proofs (e.g. of termination) but irrelevant for the execution.

Modifying the invariant does not change the program (the operational semantic) and should therefore only have limited impact on the logical semantic. The surprise is that by replacing in the previous example the invariant $m \in [0, \min(S)]$ by $m \in \mathbb{N}$, less precise but still correct, the logical semantic is radically modified. This modified logical semantic leads to a refutation of the previous proposition, that is it indicates that the substitution is not always extracting the minimum. A rather strange conclusion, as both versions of the logical semantic describe the same program.

We have also identified a similar concern with Coq. In this case there is a single language, mixing logical and computational constructs, an extraction mechanism allowing for the elimination of the former to derive from the latter a program in a functional language, e.g. in OCaml.

As already pointed out in Par. IV-A, an inductive definition such as Inductive $E : \text{Set} := \text{nxt} : E \rightarrow E$ lacks an atomic constructor and is therefore empty. Emptiness is not, by itself, inconsistent but makes possible to prove any result of the form $\forall (e : E), P$. Its extraction in OCaml is a straightforward translation to type $E = \text{nxt} \text{ of } E$. The interesting point is that this OCaml type is not empty, as it contains the value let rec $e = \text{nxt}(e)$, not valid in Coq but making possible to use a program extracted from a fully certified Coq library with unexpected (and therefore unwanted) behaviours.

It is beyond the scope of this paper to further discuss these questions, once noted that any such bias is a potential weakness usable by a malicious developer (or a trap for an honest but inattentive developer). These remarks are not intended to criticize the tremendous work represented by the
full development of the theories supporting formal methods. They however justify the interest in mechanically checking such theories, pursuing works described e.g. in [32]–[34].

B. About the tools

Beyond the concerns about the theory, one may also question the validity of the tool implementing a formal method. For example a prover can be incomplete (unable to prove results valid in the theory) or incorrect (able to prove results unprovable in the theory), the latter being more worrying, at least from the evaluation and certification perspective, as it may lead to an artificial paradox. And indeed such paradoxes have been discovered in well established tools.

Clearly, implementing a formal method is a difficult task, dealing not only with completeness, correctness, but also with performance, automation, and ergonomy. In our view, the (potential) existence of bugs in a tool does not mean that it should not be used, but that the provided results should be considered with some care, and possibly verified by other mechanisms. This is addressed for example by [29], [35].

VI. Stepping Out of the Model

We have discussed at length some concerns regarding the formal development of secure systems, through questioning paradoxes in the theory, bugs in the tools or more simply by identifying gotchas in the specifications. Let’s now assume that we have been able to produce a consistent specification with security properties correctly expressed, and a compliant implementation whose all proof obligations have been discharged, security properties correctly expressed, and a compliant implementation.

A. About Closure

A frequent implicit hypothesis is related to the use of closure proofs. For example, proving a $B$ machine requires proving the preservation of its invariant by any of its operations. This is justified if there is no other way to influence the system state than the provided operations. The extent to which this is enforced in the real system has to be carefully analysed. Threats considered during security analysis may reflect actions that are not in the model (data stored in files by proven applications can be modified by other applications, signals in electronic circuits can be jammed by fault injection, etc). There is no silver bullet to address this problem; current approaches include defensive style programming, redundancy, and dysfunctional considerations (e.g. by modelling errors such as unexpected values or inconsistent states).

B. About Typing

A second example of implicit hypothesis, much less obvious, is related to types. An adequate use of types in a specification (for example modelling IP addresses and ports as values of abstract sets rather than natural values) ensures that some forms of error will be automatically detected (such as using a port where an address is expected). But it is also important to understand how strong an hypothesis it is, and how easily it can be violated. Indeed, types are again logical information that have generally no concrete implementation; in most programming languages, they just disappear at compilation. So, while ill typed operation calls cannot be considered during formal analysis, they are in some cases executable.

A typical example is provided in [36], describing a flaw in the PKCS#11 API for cryptographic resources, summarised here. A central authority (e.g. a bank) distributes cryptographic resources to customers. Such a resource can perform cryptographic operations, $C \leftarrow cipher(M, K)$ to cipher the message $M$ with the key numbered $K$, or $M \leftarrow uncipher(C, K)$ for the inverse operation. The resource never discloses keys to the customer, but permits exchange of keys with other resources through export of wrapped (cyphered) keys using $D \leftarrow export(K, W)$ where $K$ is the number of the exported key and $W$ the number of the wrapping key, and $import(D, W, K)$ for the inverse operation (that stores internally the unwrapped key under number $K$ without disclosing it). In a model where cyphertexts and wrapped keys are of different types, one can prove that no sequence of calls will disclose a sensitive key. Unfortunately the implementations of cyphertexts and wrapped keys are indistinguishable, and stored keys are not tagged with their role. It is so possible to disclose a key $K$ with the (ill-typed) sequence $D \leftarrow export(K, W); M \leftarrow uncipher(D, W)$.

This demonstrates that it is important to identify implicit hypotheses associated to the use of types to detect possible consequences of type violations, or to maintain type information in the implementation to prevent such attacks.

VII. Conclusion

We summarise and discuss difficulties related to the development of secure systems using formal methods, identifying – where possible – proposals for improvement. The concerns described in this paper were identified during a systematic review of the process of formal development, investigating possible difficulties.

A quick read of this paper could seem to imply that the reputation of formal methods to develop correct systems is overestimated. This is not our message. We consider that formal methods are very efficient tools to obtain high level of assurance and confidence for the development of systems in general, and of secure systems in particular.

Yet to fully benefit from such tools, one has to understand their strengths but also their limitations. Pretending that proven secure systems are perfectly secure is nothing more than a renewed version of the first myth about formal methods pointed out in [37], and is to the least inadequate; in fact, we consider that such a claim is detrimental to formal methods. Taking
this into account, we expect our proposals to help, where possible, for improving the quality of formal specifications and the adequacy of formal developments of secure systems (in some cases relying on other methodologies or technologies); our second expectation is to shed some light on the difficulties to at least allow for a better evaluation of the genuine level of confidence obtained through the use of formal methods.

Note: An extended version of this paper is available in French language at [38].

REFERENCES