Variants of Validity and their Impact on the Overall Test Space

Jörg Herrmann
FIT Leipzig
at HTWK Leipzig
P.O.Box 30 00 66
04251 Leipzig, Germany
joerg@informatik.htwk-leipzig.de

Klaus P. Jantke
Meme Media Laboratory
Hokkaido University
Kita-13, Nishi-8, Kita-ku
060-862 Sapporo, Japan
jantke@meme.hokudai.ac.jp

Rainer Knauf
Technical University of Ilmenau
Faculty of Comp. Sci. & Automation
P.O.Box 10 05 65
98684 Ilmenau, Germany
Rainer.Knauf@theoinf.tu-ilmenau.de

Abstract

Dealing with the question whether or not a given system does suffice some interesting property one is confronted with the problem, how to navigate appropriately through the available knowledge space to prove or to refute that property under investigation. Moreover it is desireable to get an algorithm which solves that task efficiently.

Applied to the area of system validation, we will propose some solution allowing for the reduction of test cases, i.e. of the scale of the knowledge space to be investigated, when validating some target system by testing.

The key idea for test case reduction is to exploit certain inheritance properties of the underlying space of input data.

In its right perspective, inheritance is induction. Due to the impossibility to create any general induction scheme for deductive justification of inductive reasoning, there arises the necessity of domain-dependent variants.

Introduction

Experimenting with a system is one way to gain some understanding of its behavior. In contrast to purely observing inputs and according outputs it allows to control the process of aquiring information about the system by predefining several characteristics of the target system's environment.

For an instance of those characteristics, we might think of the order in which information is presented to the input of some inference machine (cf. (Dötsch & Jantke 1996), e.g.) Also restrictions affecting the entire set of possible input information have to be mentioned. They are subject to countless investigations in learning theory (cf. (Angluin & Smith 1983), e.g.) and do correspond to filtering of information in any domain.

But to ensure high efficiency, experiments should draw advantage not only from environmental relations, they should acknowledge also mechanisms detected to be inherent for the observed system. In that context, axiomatic knowledge concerning the system's internal structure has to be considered – we speak of so-called 'white-box' testing (cf. (Gonzalez & Ramasamy 1997), (Gupta 1993)) – as well as findings collected during the experiments.

In the following, we will restrict to a special field of experimentation – the validation of knowledge-based systems. Our interest is focused on the problem of reducing the amount of necessary test cases when validating a system by testing. Obviously, that does require some understanding of the notion *test case* and of what we call a *quasi-exhaustive test set* ((Herrmann, Jantke, & Knauf 1997b)). The reader may consult (Jantke, Abel, & Knauf 1997) for the fundamentals of our so-called TURING test approach, (Abel, Knauf, & Gonzalez 1996), (Arnold & Jantke 1997), (Gonzalez, Gupta, & Chianese 1996), and (Jantke 1997) for some applications in different areas, and, furthermore, (Jantke, Knauf, & Abel 1997), (Knauf *et al.* 1997), or (Knauf, Philippow, & Gonzalez 1997), for surveys of the comprehensive validation scenario.

Interactive approaches to system validation based on systematic experimentation typically consist of the following main phases of

- test case generation and optimization,
- experimentation by feeding in test cases,
- evaluation of experimentation results, and
- validity assessment, i.e. synthesis of some validity statement based on experimentation results,

which might be looped and dovetailed, in several ways.

We will try to control the exploration of a given system by making suggestions how to start and continue some experiment, operating about the system's structure and results already available from preceding tests.

The key technical term underlying this publication is *test case* and the key computational process is the *reduction* of sets of test cases to those subsets which are *quasi-exhaustive*. The crucial theoretical concept underlying our approach is *inheritance of validity*.

We focus our present investigation on the inheritance phenomenon for both conceptually specifying and computationally determining quasi-exhaustive test sets.

Inheritance and Induction

There has been invented some idea (for a brief survey, cf. (Herrmann 1997), e.g.) for deductively describing the relationship between complete sets of test cases and those reduced subsets which still contain sufficient information for replacing the usually unfeasibly large complete test sets during system validation.

We briefly discuss the essentials before going into more detail, in the following chapters.

Assume any target system under investigation. For any region R of input data, the formula $valid(R)$ is intended to express the validity of the given system on all data belonging to R. If R is large, a pointwise testing of $valid(s)$, for all $s \in R$, might be computationally unfeasible. It is highly desirable to find some fairly small subset S of R such that (i) $valid(S)$ can be experimentally justified and, furthermore, (ii) there is a logical relationship $rel(S,R)$ which justifies the conclusion of $valid(R)$ from $valid(S)$, only. Formally,

$$valid(S) \wedge rel(S,R) \longrightarrow valid(R)$$

should be logically valid. In its right perspective, this formula is representing some scheme of induction.

Even more general, the desired formula is parameterized by the target predicate *valid*. In (Shoham 1988), the author develops several types of inheritance which might apply or not to certain predicates within some particularly given logical framework. The relationship *rel* which allows for the inheritance of some predicate's validity depends, in general, on the specific predicate *pred* under consideration. Thus, the syntactic version

$$pred(S) \wedge rel_{pred}(S,R) \longrightarrow pred(R)$$

is more appropriate. This is setting the stage for a clear formulation of our present investigation's aim:

Given some target system and some related concept of validity, which is formalized over input data by some predicate *valid*, find any logical formula $rel_{valid}(S,R)$ such that the corresponding instance of the induction scheme displayed above becomes logically true, i.e.

$$valid(S) \wedge rel_{valid}(S,R) \longrightarrow valid(R)$$

constitutes a theorem in its own right.

When such a formula rel_{valid} has been found, the practically important question is how to exploit this knowledge algorithmically: Given any data set R with

- $valid(R)$

describing some domain of interest within the target behaviour, construct any data set S, which is usually assumed to be a subset of R, such that

- $valid(S)$ and
- $rel_{valid}(S,R)$

are satisfied.

From Karl Popper's seminal work (cf. (Popper 1934) resp. (Popper 1965)), it is already clear that there does not exist any universal approach towards a deductive justification of induction. Consequently, one has to search for domain-specific and even for system-specific variants of validity and its inheritance.

An Introductory Illustration

The present section is intended to exemplify the rather abstract concepts developed within the preceding one. Our focus is threefold. First, the underlying ideas shall be lucidly illustrated, for the reader's convenience. In doing so for only a toy example, the reader might get an impression of the difficulties faced under more realistic conditions. Second, the example considered will indicate the necessity to process knowledge about the target behaviour and about the system under inspection, thus relating black box validation to white box validation from the perspective of test case reduction. Third, even the almost trivial toy example will allow for first steps towards the distinction of several logical solutions. This might be understood a supplementary motivation of our present investigation.

Assume we deal with classification problems. For simplicity, the underlying space of input data is only one-dimensional. All variables which occur below are real-valued. A generalization to higher dimensions is straightforward. To cap it all, we confine ourselves to discussing the system's behaviour over a single interval, only.

If a, b', and b are any three reals satisfying $a < b' < b$, one might consider the following two formulae:

$$\begin{aligned} rel^{(i)}_{valid}(S,R) \quad \equiv \quad & R = [a,b) \wedge \\ & (b',b) \subseteq S \wedge a \in S \wedge \\ & \forall x \, (x \in [a,b'] \longrightarrow \\ & \qquad f(a) = f(x) \vee f(x) = f(b') \,) \\ rel^{(ii)}_{valid}(S,R) \quad \equiv \quad & R = [a,b) \wedge \\ & \{a,b'\} \subseteq S \wedge \\ & \forall x \, (x \in (b',b) \rightarrow f(b') = f(x) \,) \wedge \\ & \forall x \, (x \in [a,b'] \longrightarrow \\ & \qquad f(a) = f(x) \vee f(x) = f(b') \,) \end{aligned}$$

In these two formulae, open, semi-open, and closed intervals are denoted as usual. The function symbol f is chosen to represent the classifier's input/output behaviour, i.e. $f(x)$ names the classification result for the input x, e.g.

Note that our investigations need to be based on some concept of validity of logical formulae which is not sufficiently clear, yet. In particular, it is necessary to specify where f refers to the target behaviour and where it does reflect the actually experienced system's behaviour. We will return to this crucial point in a later section.

Before going into these details, let us investigate how to invoke these formulae for test set reduction. For simplicity, assume that a *valid* classifier should return, for every element of the domain R, a certain fixed class name c. Formally, for any subset $X \subseteq R$, $valid(X)$ holds if and only if the formula $\forall x \in X \, (f(x) = c)$ is true. Clearly, there is no hope to test all points in any given interval $R = [a,b)$.

Given any interval $R = [a,b)$, utilize $rel^{(i)}_{valid}(S,R)$ resp. $rel^{(ii)}_{valid}(S,R)$ to find a suitably small set $S \subseteq R$ with $valid(S)$. By inheritance, this will guarantee the

system's validity over R. In formal terms, reduction of test sets is justified by the two *Theorems*

$$valid(S) \wedge rel^{(i)}_{valid}(S,R) \longrightarrow valid(R)$$

$$valid(S) \wedge rel^{(ii)}_{valid}(S,R) \longrightarrow valid(R)$$

which might be easy to prove (see below).

The reader should be aware of the key differences between the two formulae under consideration which might be expressed sufficiently clear as lemmata, e.g. The operator # will be used to indicate any given set's cardinality.

Lemma 1 *For any interval $R=[a,b)$ and for any subset $S \subseteq R$, it holds $rel^{(i)}_{valid}(S,R) \longrightarrow \#(S)=\infty$.*

Lemma 2 *For any interval $R=[a,b)$, there are subsets $S \subseteq R$ with $rel^{(ii)}_{valid}(S,R)$ and $\#(S)<\infty$.*

Although we deal only with some toy example, the distinction clarified by the two lemmata above bears abundant evidence for the necessity to find some appropriate inheritance concepts for test set reduction. The formula $rel^{(ii)}_{valid}(S,R)$ seems more suitable than $rel^{(i)}_{valid}(S,R)$.

We confine ourselves to a brief discussion of one more problem in some detail: the knowledge underlying the justified reduction of test sets.

Parts of the formulae $rel^{(i)}_{valid}(S,R)$ and $rel^{(ii)}_{valid}(S,R)$ are determined by the target behaviour. This applies to the choice of the bounding values a and b, in particular. However, proving the "theorem"

$$S \subseteq R \wedge valid(S) \wedge rel^{(ii)}_{valid}(S,R) \longrightarrow valid(R)$$

requires knowledge about the given system. Consider the choice of an appropriate value b'. To prove the implication $\forall x(x \in (b',b) \rightarrow f(b')=f(x))$, knowledge about the system under investigation is inevitable. In the case of *white box validation*, the necessary knowledge might be easily available. Otherwise, one is facing another validation problem.

We conclude this section by some short summary from the viewpoint of inheritance and induction. In their right perspective, the formulae introduced above determine some induction principle. Obviously, there is some tradeoff:

- Based on a stronger induction principle rel_{valid}, one may be able to arrive at weaker preconditions for justification of $valid(R)$, i.e. one may rely on smaller sets S of test cases.
- The justification of stronger induction principles will usually require the investment of more expressive knowledge. It might depend on specific knowledge about the system under investigation, in particular.

In fact, all consideration above including theorems and lemmata assume some formal validity concept to be supplemented below.

Quasi-Exhaustive Test Sets

So far, we did refrain from an in-depth discussion of technical concepts like "test case", for instance. As a prerequisite of successive search for relations like $pred(S)$ or $rel_{pred}(S,R)$, we point on several possible definitions. The technicalities will be illustrated.

Test Case Concepts

Since the goal of testing in any domain might be understood as gathering new information about some object of interest, but less as translation of existing data into another format only, it makes sense to start from considering the testing of "black boxes" instead of so called "white boxes" (cf. (Gupta 1993), e.g.). The conceptual difference between both "black" and "white" boxes consists in the absence of knowledge concerning the functional behaviour or structural peculiarities of the first one, permitting to conclude on the object's reactions in case of external perturbations.

Though we are often confronted with systems of an intermediate type, it is desireable to define the notion "black box" not too strong. In particular, testing leads – except for some trivial case – to sequences of information that do not transform a black box immediately into a white one. Rather, they do enlighten the system's characteristics gradually.

Hence, suppose some black box the behaviour of which is widely unknown and that interacts with its environment exclusively via junctions X and Y, called inputs resp. outputs. Both spaces might be deeply structured. However, we may suppress those details throughout the present paper. The range of values accepted for X is denoted by X, the output space is Y.

Test cases are finite respectively finitely describable[1] data collections expressing some potential input/output behaviour. The space of all potential test cases is:

$$T = X \times (2^Y \setminus \emptyset)$$

where, for every test case (x,y), x is called *test data*.

In dependence of the kind of system we deal with, there do exist several ways for extending this approach.

For instance, one might consider sequences of those test cases t (possibly of some fixed maximum length n), thus expanding the test space T:

$$\forall i<n:\ T^{i+1} = \{\, t \circ t^i \mid t \in T = T^1 \ \wedge\ t^i \in T^i \,\} \cup T^i$$

According extensions can be useful when investigating systems that store the order of information accepted from their inputs (cf. (Böhme 1995)), i.e. for *interactive dynamic* systems. On the other hand, there are extrema like autonomous automata that do not take any notice of possible inputs, without losing the ability to show an unexpected resp. non-deterministic behaviour.

[1]The reader should recall that intervals of rationals or of reals, e.g., are usually finitely describable, although they are conceptually infinite.

Variants of Validity Concepts

Very roughly, but also very intuitively, validation as understood in the present investigation deals with the derivation of validaty assessments through interactive experimentation. Complex validity assessments are built upon elementary statements which express the validity of the present system over certain domains of interest. Thus, elementary goals to arrive at are certain formulae which formalize desirable system properties that establish the quality of a system's validity.

The inspected *system's validity* is the ultimate goal of investigations, but the technical results arrived at are formulae. A certain *formula's validity* carries part of the information about the system's quality.

For this purpose, one has to determine the meaning of a given formula being valid with respect to a given system's behaviour. For the particular purpose of test set reduction, the target behaviour comes into play, as briefly illustrated above. Consequently, a formula's validity might refer to a given system's behaviour, to an implicitly given target behaviour, or even to both of them. This is far from being standard and, thus, requires some careful specification. A sketch must do.

From a most general and sufficiently formal point of view, any input/output behaviour may be abstracted as a relation over $X \times Y$. For representing knowledge about the behaviour, one needs a relational symbol, say f (which was assumed to be functional, for simplicity, in our introductory example).

We assume any standard language of first order predicate calculus extended by the extra symbol f.

There must exist any uniform agreement about the underlying validity concept for arbitrary formulae φ. For the sake of this short presentation, the reader might assume any standard technique of interpreting formulae including arithmetics, e.g. But what about non-standard formulae in which f occurs?

Recall that there is a necessity to interpret f either with respect to the target behaviour or with respect to the current system under inspection.

In case of system *verification*, one might assume some specification of f. In logical terms, this would establish a theory. Then, validity were straightforward.

But when dealing with system *validation*, there is usually no theory accessible which describes the target behaviour sufficiently complete. Knowledge is episodic and incomplete, and presented ad hoc. The results of experimentation are quite similar in character. After a finite number of experiments with the current system, there is a finite number of so-called protocols which incompletely describe the system's actual behaviour. Furthermore, knowledge is usually uncertain, due to tolerances of sensors and limited precision of available instruments like clocks, e.g.

From a formal perspective, both the knowledge about the desired behaviour and the knowledge which is resulting from exploratory experimentation is forming finite subsets of models. Every such finite subset F implicitly describes the class of all its potential completions $\mathcal{M} = \{M \mid F \subseteq M\}$. When time proceeds, subsequent experimentations lead to enrichments of the available knowledge. We are dealing with sequences of models $\{F_n\}_{n=1,2,\ldots}$ implicitly specifying model classes $\mathcal{M}_n = \{M \mid F_n \subseteq M_n\}$. The sequence $\mathcal{M}_n$ is shrinking over time, thus reflecting an increasing degree of precision. Note that this is essentially the approach underlying (Arnold 1994) and (Arnold & Jantke 1996).

At a first glance, there seems to be an easy way out. If any given F_n represents the currently available knowledge, the validity of a particular formula φ might be understood as $\mathcal{M}_n \models \varphi$. But this does not do.

Usually, there are several ways to specify the validity of formulae over incomplete and uncertain knowledge.

Assume, for illustration, a relational classifier which is expected to meet on some particular input x_0 the condition

$$f(x_0) \geq high$$

and, furthermore, that some experimentation during interactive validation yields

$$f(x_0) \subseteq \{medium, high\}$$

about the system's actual behaviour. Does this meet the requirement $f(x_0) \geq high$? Just for another quite similar illustration, the reader may imagine the value of a real-valued, but imprecisely measured function to be represented as $f(x_1) = [0.1, 0.4]$. What about the condition $f(x_1) < 0.2$? Is this formula valid or not?

There are at least two obvious extremes of specifying validity: either calling a formula valid if and only if it is classically true under every restriction of the originally given non-determinism or calling it valid if there is at least one appropriate constellation of data. These variants are well-known in modal logics and result in different expressive power. The corresponding notions are *necessity* resp. *possibility*.

Quasi–Exhaustiveness

Test sets may depend on the system's factual behaviour and on its intended behaviour, as well. For combination of both parameters, we have introduced the concept of quasi-exhaustiveness by defining relations $qexh_{valid}(S,R)$. For subspaces of test data $R \subseteq X$, one needs to derive formulae

$$valid(S) \wedge qexh_{valid}(S,R) \longrightarrow valid(R)$$

where all three subformulae usually contain the extra relational symbol f introduced above. Thus, non-standard concepts of validity have to be invoked.

For the quality of an example, figure 1 illustrates some 2-dimensional test space, used for validation of any underlying system. Shaded areas within that test space do cover so-called *regions of inheritance*. For each of these regions specific inheritance mechanisms w.r.t. the system's validity are supposed to be known. One can distinguish 5 areas R_i, denoted by $i \in \{a, \ldots, e\}$. The R_i can differ concerning the type of inheritance as

well as the type of validity required over the according region. In its consequence, one may follow several validation scenarios for which we only sketch two extrema.

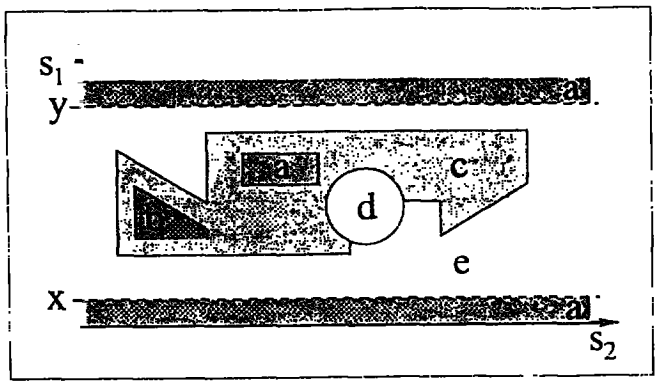

Figure 1: Regions of Inheritance for a 2-dimensional Test Space

Figure 2 shows several minimal quasi-exhaustive test sets for the mentioned regions R_i. Here, the system's validity is inherited uniformly from some environment of the region corners to region boundaries, and from boundaries to the according regions (cf. (Abel, Knauf, & Gonzalez 1996), (Abel & Gonzalez 1997b), etc.).

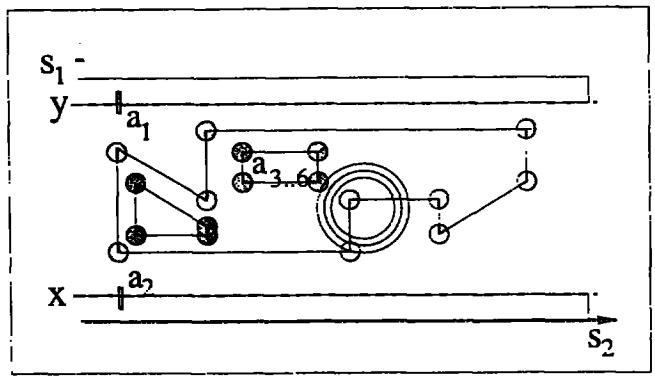

Figure 2: Quasi-exhaustive Test Sets

A second criterion, not presented in the figure, might consider the limited resolution of any measurement, resulting in some discreteness of possible sensor values ((Arnold 1994), (Herrmann, Jantke, & Knauf 1997a)).

Third, we did acknowledge the assignment of inputs that do not affect the output behaviour currently under investigation. This assignment can be resolved arbitrarily, producing in figure 2 the union of six sets $a_1 \ldots a_6$ in the quality of some quasi-exhaustive test.

There do exist more scenarios for reducing test sets, leading to different kinds of validity or different costs in testing and evaluation of test results (cf. (Abel & Gonzalez 1997b), (Herrmann 1997), (Abel & Gonzalez 1997a)). However, exploration of appropriate inheritance mechanisms presents a large field for future work.

Generating Minimal Test Sets

Some operational idea for constructing minimal quasi-exhaustive test sets was developed in (Herrmann, Jantke, & Knauf 1997a). Now, we exemplify a first refinement. Like the basic approach, it starts from any default test space D that will be reduced by combining certain regions of inheritance:

begin
  $TestSpace := D$
  for any R
    $R \cap TestSpace \neq \emptyset$
    $\wedge\ \exists Q\colon qexh^{min}_{valid}(Q,R)$
  do
    $TestSpace := TestSpace \backslash (R \cap \overline{Q}) \cup (Q \cap \overline{D})$
  enddo
  return $TestSpace$
end

As long as there do exist regions of inheritance w.r.t. *valid*, the current property of interest, these regions are applied to the test space, reducing it trough substitution by some quasi-exhaustive subset. To avoid repeated reductions of the same region, application is limited to minimal quasi-exhaustive test sets. Similar, we avoid to include subsets of D that have been excluded during preceding steps.

Additionally, one could forbid iterations with

$$\|R \cap \overline{Q}\| \leq \|Q \cap \overline{D}\|$$

since they might increase the cardinality of $TestSpace$ instead of decreasing it. However, there are domains where similar hill climbing techniques will fail.

The algorithm presented above is inherently nondeterministic. In contrast, the problem on which we focus is order dependent und resembles the Travelling Salesman Problem which has been shown to be NP-complete (Garey & Johnson 1979). Depending on the strategy for selecting some *region of inheritance* in each iteration, we gain reduced test sets of different cardinality or composition. This raises the question which criteria to apply for selecting a "best" region in each reduction step or in certain of them, subsequently.

We adopt Robin Murphy's view in her talk on FLAIRS'97 by looking for ways "... to solve some task only as precise as necessary", to make light of the combinatorial problems we are facing. In fact, one has to ask simultaneously for criteria signaling that further iterations do not lead to a considerable improvement of some intermediate result. One can express this problem as follows: When will costs for reducing a test set reach or exceed the costs for testing it?

However, the task of finding an optimal way for overlapping regions is not yet solved. So we proceed to look for feasible heuristics, drawn from graph theory. In particular, there remains the question how to combine optimal ways that contain identical sections?

References

Abel, T., and Gonzalez, A. J. 1997a. Influences of criteria on the validation of Ai systems. In Wittig, W. S., and Grieser, G., eds., *LIT'97, 5. Leipziger Informatik-Tage an der HTWK Leipzig, 25-26. September 1997, Tagungsbericht*, 83–88. Forschungsinstitut für InformationsTechnologien Leipzig e.V. (FIT Leipzig e.V.).

Abel, T., and Gonzalez, A. J. 1997b. Utilizing criteria to reduce a set of test cases for expert system validation. In Dankel II, D. D., ed., *FLAIRS-97, Proc. Florida AI Research Symposium, Daytona Beach, FL, USA, May 11-14, 1997*, 402–406. Florida AI Research Society.

Abel, T.; Knauf, R.; and Gonzalez, A. 1996. Generation of a minimal set of test cases that is functionally equivalent to an exhaustive set, for use in knowledge-based system validation. In Stewman, J. H., ed., *Proc. Florida AI Research Symposium (FLAIRS-96), Key West, FL, USA, May 20-22, 1996*, 280–284. Florida AI Research Society.

Angluin, D., and Smith, C. H. 1983. A survey of inductive inference: Theory and methods. *Computing Surveys* 15:237–269.

Arnold, O., and Jantke, K. P. 1996. Representing and processing dynamic knowledge in complex dynamic systems. In Alpaydın, E.; Çilingiroğlu, U.; Gürgen, F.; and Güzeliş, C., eds., *Fifth Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN'96, Istanbul, June 27-28, 1996*, 351–356. Boğaziçi University.

Arnold, O., and Jantke, K. P. 1997. Towards validation of Internet agents. In Wittig, W. S., and Grieser, G., eds., *LIT-97, Proc. 5. Leipziger Informatik-Tage, Leipzig, 25./26. September 1997*, 89–100. Forschungsinstitut für InformationsTechnologien Leipzig e.V.

Arnold, O. 1994. A logic of constraints for dynamic process control. WISCON Report 09/94, HTWK Leipzig (FH), Fachbereich IMN.

Böhme, R. 1995. *Ein Ansatz zur wissensbasierten Störungsanalyse verfahrenstechnischer Systeme.* Ph.D. Dissertation, TH Leipzig.

Dötsch, V., and Jantke, K. P. 1996. Solving stabilization problems in case-based knowledge acquisition. In Compton, P.; Mizoguchi, R.; Motoda, H.; and Menzies, T., eds., *Pacific Knowledge Acquisition Workshop, Oktober 23-25, 1996, Sydney, Australia*, 150–169. University of New South Wales, Department of Artificial Intelligence.

Garey, M. R., and Johnson, D. S. 1979. *Computers and Intractability.* W.H. Freeman and Co.

Gonzalez, A. J., and Ramasamy, P. 1997. Detecting anomalies in constraint-based systems. In Gens, W., ed., *IWK-97, 42nd International Scientific Colloquium, Ilmenau University of Technology*, volume 2, 35–40. TU Ilmenau.

Gonzalez, A. J.; Gupta, U. G.; and Chianese, R. B. 1996. Performance evaluation of a large diagnostic expert system using a heuristic test case generator. *Engineering Applications of Artificial Intelligence* 1(3):275–284.

Gupta, U. G. 1993. *Validation and Verification of Expert Systems.* IEEE Press, Los Alamitos, CA.

Herrmann, J.; Jantke, K. P.; and Knauf, R. 1997a. Towards cost driven system validation. In Gens, W., ed., *IWK-97, 42nd International Scientific Colloquium, Ilmenau University of Technology*, volume 2, 41–46. TU Ilmenau.

Herrmann, J.; Jantke, K. P.; and Knauf, R. 1997b. Using structural knowledge for system validation. In Dankel II, D. D., ed., *FLAIRS-97, Proc. Florida AI Research Symposium, Daytona Beach, FL, USA, May 11-14, 1997*, 82–86. Florida AI Research Society.

Herrmann, J. 1997. Order dependency of principles for reducing test sets. In Wittig, W. S., and Grieser, G., eds., *LIT'97, 5. Leipziger Informatik-Tage an der HTWK Leipzig, 25-26. September 1997, Tagungsbericht*, 77–82. Forschungsinstitut für InformationsTechnologien Leipzig e.V. (FIT Leipzig e.V.).

Jantke, K. P.; Abel, T.; and Knauf, R. 1997. Fundamentals of a TURING test approach to validation. Technical Report 01/97, Hokkaido University Sapporo, Meme Media Laboratory.

Jantke, K. P.; Knauf, R.; and Abel, T. 1997. The TURING test approach to validation. In Terano, T., ed., *15th International Joint Conference on Artificial Intelligence, IJCAI-97, Workshop W32, Validation, Verification & Refinement of AI Systems & Subsystems, August 1997, Nagoya, Japan*, 35–45.

Jantke, K. P. 1997. Towards validation of data mining systems. In Wittig, W. S., and Grieser, G., eds., *LIT-97, Proc. 5. Leipziger Informatik-Tage, Leipzig, 25./26. September 1997*, 101–109. Forschungsinstitut für InformationsTechnologien Leipzig e.V.

Knauf, R.; Jantke, K. P.; Abel, T.; and Philippow, I. 1997. Fundamentals of a TURING test approach to validation of AI systems. In Gens, W., ed., *IWK-97, 42nd International Scientific Colloquium, Ilmenau University of Technology*, volume 2, 59–64. TU Ilmenau.

Knauf, R.; Philippow, I.; and Gonzalez, A. J. 1997. Towards an assessment of an AI system's validity by a Turing test. In Dankel II, D. D., ed., *FLAIRS-97, Proc. Florida AI Research Symposium, Daytona Beach, FL, USA, May 11-14, 1997*, 397–401. Florida AI Research Society.

Popper, K. 1934. *Logik der Forschung.* J.C.B. Mohr, Tübingen.

Popper, K. 1965. *The Logic of Scientific Discovery.* Harper & Row.

Shoham, Y. 1988. *Reasoning about Change.* Cambridge, MA: MIT Press.