Chapter 1
Fault Tolerance and Resilience: Meanings, Measures and Assessment

Lorenzo Strigini

Abstract To assess in quantitative terms the "resilience" of systems, it is necessary to ask first what is meant by "resilience", whether it is a single attribute or several, which measure or measures appropriately characterise it. This chapter covers: the technical meanings that the word "resilience" has assumed, and its role in the debates about how best to achieve reliability, safety, etc.; the different possible measures for the attributes that the word designates, with their different pros and cons in terms of ease of empirical assessment and suitability for supporting prediction and decision making; the similarity between these concepts, measures and attached problems in various fields of engineering, and how lessons can be propagated between them.

1.1 Introduction

Measuring or assessing a quality for any object, e.g., "resilience" for a system, requires clarity about what this quality is.

The word "resilience" has become popular in recent years in the area of information and communication technology (ICT) and policy related to ICT, as part of a more general trend (for instance, the word "resilience" is in favour in the area of critical infrastructure protection). The increasing use of this word creates the doubt whether it is just a new linguistic fashion, for referring to what is commonly studied, pursued and assessed under names like "fault tolerance", "dependability" (a term mostly restricted to ICT usage), "security", "reliability, availability, maintainability and safety" (RAMS), "human reliability", and so on, or it actually denotes new concepts. While there may be a component of fashion, the increased use of the word

L. Strigini (✉)
Centre for Software Reliability,
City University London,
Northampton Square, London EC1V 0HB, UK
e-mail: strigini@csr.city.ac.uk

K. Wolter et al. (eds.), *Resilience Assessment and Evaluation of Computing Systems*,
DOI: 10.1007/978-3-642-29032-9_1, © Springer-Verlag Berlin Heidelberg 2012

"resilience" is often meant to highlight either novel attention to these problems or a plea for a shift of focus in addressing them. It is useful to consider what these new foci may be and whether they require new concepts and new measures. Technical, and especially quantitative, reasoning about "resilience" requires clear definitions of these concepts, whether old or new.

Without reviewing in detail the multiple uses of "resilience", it is useful to recognise how the technical problems and debates in which it appears in different areas of application are related, highlighting similarities and differences in the problems they pose for quantitative reasoning, including measurement and benchmarking, and retrospective assessment as well as prediction.

The word "resilience", from the Latin verb *resilire* (*re-salire*: to jump back), means literally the tendency or ability to spring back, and thus the ability of a body to recover its normal size and shape after being pushed or pulled out of shape, and therefore figuratively any ability to recover to normality after a disturbance. Thus the word is used technically with reference to materials recovering elastically after being compressed, and also in a variety of disciplines to designate properties related to being able to withstand shocks and deviations from the intended state and go back to a pre-existing, or a desirable or acceptable, state. Other engineering concepts that are related to resilience therefore include for instance fault tolerance, redundancy, stability, feedback control.

A review of scientific uses of the word "resilience" for the European project *ReSIST* ("Resilience for Survivability in IST") [770] identified uses in child psychology and psychiatry, ecology, business and industrial safety. In many cases, this word is used with its general, everyday meaning. Some users, however, adopt specialised meanings, to use "resilience" as a technical term.

The premise for calling for an everyday word to be used with a new specialised meaning is that there is a concept that needs to have its own name, for convenience of communication, and lacks one. The concept is sometimes a new one ("entropy", for instance), or a new refinement of old concepts ("energy", for instance), or just a concept that needs to be referred to more often than previously (because the problems to be discussed have evolved) and thus requires a specialised word. Sometimes, the motivation is that words previously used for the same concept have been commandeered to denote, in a certain technical community, a more restricted meaning: for instance, after the word "reliability" acquired a technical meaning that was much more restrictive than its everyday meaning, the word "dependability" came to be used, by parts of the ICT technical community, to denote the everyday meaning of "reliability" [63].

For the word "resilience", a tendency has been to use it, in each specific community, to indicate a more flexible, more dynamic and/or less prescriptive approach to achieving dependability, compared to common practices in that community. Thus the above-cited document [770], for instance, concluded that a useful meaning to apply to "resilience" for current and future ICT is "ability to deliver, maintain, improve service when facing threats and evolutionary changes": that is, the important extension to emphasise in comparison with words like "fault tolerance" was the fact that the perturbations that current and future systems have to tolerate include change. While

existing practices of dependable design deal reasonably well with achieving and predicting dependability in ICT systems that are relatively closed and unchanging, the tendency to making all kinds of ICT systems more interconnected, open, and able to change without new intervention by designers, is making existing techniques inadequate to deliver the same levels of dependability. For instance, evolution itself of the system and its uses impairs dependability: new components "create" system design faults or vulnerabilities by feature interaction or by triggering pre-existing bugs in existing components; likewise, new patterns of use arise, new interconnections open the system to attack by new potential adversaries, and so on [769].

For a comparison with another field of engineering, a document on "infrastructure resilience" [632] identifies "resilience" as an extension of "protection". As an example of the direction for this extension, this paper questions whether burying the cables of a power distribution grid to prevent hurricane damage is "resilience", but suggests that installing redundant cabling is.

An important specialised use of the word "resilience" has emerged with "resilience engineering", a movement, or a new sub-discipline, in the area of safety (or, more generally, performance under extreme conditions) of complex socio-technical systems. Here, the word "resilience" is used to identify enhanced ability to deal with the unexpected, or a more flexible approach to achieving safety than the current mainstream approaches. The meaning is somewhat different between authors, which need not cause confusion if we consider "resilience engineering", rather than "resilience", as the focal concept for these researchers, and actually a neologism, designating an area of studies and the ongoing debate about it. This area will be further discussed below.

From the viewpoint of the problems of quantitative assessment, measurement and benchmarking, the goals of these activities and the difficulties they present, there is no sharp boundary between the socio-technical systems that are of concern to ICT specialists and those addressed by "resilience engineering". There are undoubtedly differences in the typical scales of the systems considered, but the progress in ICT towards the "future Internet" and greater interconnection of ICT with other infrastructures and activities are cancelling these differences [769]. Most dependability problems in ICT have always involved some social and human factors influencing dependability, for instance through design methods and constraints, or through the maintenance or use of technical systems. In this sense, ICT dependability is about socio-technical systems. As ICT becomes more pervasive and interlaced with human activities, the dependability of the technical components in isolation may become a minor part of the necessary study of dependability and thus of resilience. For example, this occurs in a hospital or air traffic control system, where automated and human tasks interact, and contribute redundancy for each other, on a fine-grain scale. It also occurs where large scale systems involve networks of responsibilities across multiple organisations, as in the provision of services (possibly through open, dynamic collaboration) on the present or future Internet.

In view of these similarities and disappearing boundaries between different categories of systems, this short survey, written from the vantage point of practices in the technical side of ICT dependability assessment, tries to emphasise the possible

new problems, or desirable new viewpoints, that may come from the progressive extension of the domain that ICT specialists have to study towards systems with a more important and more complex social component.

1.2 The "Resilience Engineering" Movement

The title "resilience engineering" has been adopted recently by a movement, or emerging discipline or community, started around a set of safety experts dealing mostly with complex socio-technical systems, like for instance industrial plant, railways, hospitals. A few symposia have taken place focusing on this topic and books have been published. This movement uses the term "resilience engineering" to designate "a new way of thinking about safety" [767]. The focus of these researchers is on moving beyond limitations they see in the now-established forms of the pursuit of safety: too much focus on identifying all possible mechanisms leading to accidents and providing pre-planned defences against them; too little attention to the potential of people for responding to deviations from desirable states and behaviours of the system. Thus the resilience engineering authors underscore the needs for reactivity and flexibility, e.g., "The traits of resilience include experience, intuition, improvisation, expecting the unexpected, examining preconceptions, thinking outside the box, and taking advantage of fortuitous events. Each trait is complementary, and each has the character of a double-edged sword" [681].

In using the term "resilience", there is a range between authors focusing on the resilient *behaviour* of the socio-technical system—its visibly rebounding from deviations and returning to (or continuing in) a desirable way of functioning—and those who focus on the characteristics they believe the system must have in order to exhibit such behaviour, like for instance the cultural characteristics and attitudes in the above quote. This degree of ambiguity need not cause confusion if we simply use the "resilience engineering" phrase to designate a set of related concerns, rather than "resilience" as a specific technical term. It points, however, at the variety of attributes—whatever we may call them—that are inevitably of interest to measure or predict.

Importantly, authors in "resilience engineering" underscore the difference between "resilience" and "safety", the former being just one of the possible means to achieve the latter. Their concern is often one of balance, as they see excessive emphasis on (and perhaps complacency about the effectiveness of) static means for achieving safety, designed in response to accidents, while they see a need for a culture of self-awareness, learning how things really work in the organisation (real processes may be very different from the designed, "official" procedures), taking advantage of the workers' resourcefulness and experience in dealing with anomalies, paying attention to the potential for unforeseen risks, fostering fresh views and criticism of an organisation's own model of risk, and so on. On the other hand, safety can be achieved in organisations that do not depend on "resilience" in this sense of the word, but on rigid, pre-designed and hierarchical approaches [405].

1.3 The Appeal of Resilience and Fault Tolerance

Before discussing issues of measurement and quantitative assessment, it is useful to identify some concepts and historical changes that are common to the various technical fields we consider.

When something is required to operate dependably (in a general sense, including "being secure against intentional harm"), the means available for ensuring this dependability include mixes of what in the ICT world are often called "fault avoidance" and "fault tolerance" [63]. The former means making components (including, by a stretch of the word "component", the design of the system, with its potential defects that may cause failures of the system) less likely to contain or develop faults, the latter means making the system able to tolerate the effects of these faults.

1.3.1 Historical Shifts Between "Fault Avoidance" and "Fault Tolerance"

Historically, the balance between the two approaches is subject to shifts, as is the level of system aggregation at which fault tolerance is applied. For instance, to protect the services delivered by a computer, a designer may add inside the computer redundant component(s) to form a fault-tolerant computer. Alternatively, the designer of a system using the computer (say, an automated assembly line) might provide a rapid repair service, or stand-by computers to be switched in by manual intervention, or manual controls for operators to take control if the computer fails: all these latter provisions make the control function of the assembly line fault-tolerant (to different degrees). This is a case of shift from fault tolerance in the architecture of a system component (the computer) to fault tolerance in the architecture of the system (the assembly line).

Fault tolerance (for various purposes, e.g., masking permanently disabled components, preventing especially severe effects of failures,[1] recovering from undesired transients) is a normal feature of much engineering design as well as organisation design. Fault tolerance against some computer-caused problems is nowadays a normal feature within computer architecture, but over time, as computers in an organisation or engineered plant become more numerous, the space for forms of fault tolerance "outside the computer" increased. Much of the computer hardware and software is obtained off-the-shelf, meaning that for the organisation achieving great confidence in their dependability may be unfeasible or expensive, but on the other hand there is a choice of alternatives for error confinement and degraded or reconfigured operation (relying on mixes of people and technology) if only some of these

[1] Including "system design failures": all components function as specified, but it turns out that in the specific circumstances the combination of these specified behaviours ends in system failure: the system's *design* was "faulty".

components fail, and for selectively deploying redundant automation (or redundant people) where appropriate.

Shifts of balance between fault tolerance and fault avoidance, and across levels of application of fault tolerance, occur over time with changes in technology, system size and requirements. Shifts away from fault tolerance are naturally motivated by components becoming more dependable, or their failure behaviour better known (so that fault tolerance is revealed to be overkill), or the system dependability requirements becoming (or being recognised to be) less stringent. Shifts towards more fault tolerance are often due to the observation that fault avoidance does not seem to deliver sufficient dependability, or has reached a point of diminishing returns, and in particular that good fault tolerance will tolerate a variety of different anomalous situation and faults, including unexpected ones. Thus, fault tolerance for instance often proves to be an effective defence against faults that the designers of components do not know to be possible and thus would not have attempted to avoid.

Examples of these factors recur in the history of computing, and can be traced to some extent through the arguments presented at the time to argue that the state of technology and application demanded a shift of emphasis: for instance in the papers by Avizienis in the 1970s [61] arguing for a return to more fault tolerance in computers; those of the "Recovery Oriented Computing" project in the early years of the twenty-first century [109] arguing for attention to more dynamic fault tolerance, in systems comprising multiple computers and operators. In the area of security, similar reasons motivated arguments for more of a "fault tolerance" oriented approach [302], later reinforced by concerns about the inevitable use of off-the-shelf computers and operating systems [63]. Similar considerations have applied to the proposals for fault tolerance against software faults [191, 740]. More recently, a call for papers on "Resiliency in High Performance Computing" [768] points at how the scaling up of massively parallel computations implies that the likelihood of at least one component failing during the computation has become too high if the computation is not able to tolerate such failures; similar considerations have arisen for the number of components in chips, or networks, etc, repeatedly over the years. For an example in larger systems that go beyond ICT, we may consider titles like "Moving from Infrastructure Protection to Infrastructure Resilience" [383], advocating a shift from a perceived over-emphasis on blocking threats before they affect critical infrastructure (e.g., electrical distribution grids) to making the latter better able to react to disruption. All these arguments must rely implicitly on some quantification of the risk involved by each alternative defensive solution—a sound argument about which solution entails the least risk, even without giving explicit numerical risk estimates for the individual solutions—although this quantification is not very visible in the literature.

1.3.2 Evolving versus Unchanging Redundancy

A related, recurrent line of debate is that advocating more flexible and powerful fault tolerance, in which fault tolerance mechanisms, rather than following narrowly

pre-defined strategies, can react autonomously and even evolve in response to new situations, like the human mind or perhaps the human immune system [15, 62]. Some of the recent "autonomic computing" literature echoes these themes [454]. The trade-off here is that one may have to accept a risk that the fault-tolerant mechanisms themselves will exhibit, due to their flexibility and complexity, unforeseen and sometime harmful behaviour, in return for an expectation of better ability to deal with variable, imperfectly known and evolving threats. The challenge is to assess this balance of risks, and to what extent a sound quantitative approach is feasible.

In the social sciences' approach to these problems, observations about the importance of redundancy and flexibility underpin the literature about "high reliability organisations" [783] and to some extent about "safety cultures". In this picture, the "resilience engineering" movement could be seen as just another shift in which dynamic reaction (fault tolerance) to anomalies is seen as preferable to prior provisions against them, as a precaution against unexpected anomalies. Its claim to novelty with respect to the community where it originated is in part a focus on the importance of the unexpected. This summary of course does not do justice to the wealth of specific competence about safety in organisations in the "resilience engineering" literature, or about computer failure, human error, distribution networks etc to be found in the other specialised literature mentioned above. Our goal here is to identify broad similarities and differences and their implications on assessment, measuring and benchmarking.

Much current emphasis in "resilience engineering" is about flexibility of people and organisations, not just in reacting to individual incidents and anomalous situations, but also in learning from them and thus developing an ability to react to the set of problems concretely occurring in operation, even if not anticipated by designers of the machinery or of the organisation. There is for instance an emphasis, marking recent evolution in the "human factors" literature, on the importance of understanding work practices as they are, as opposed as to how they have been designed to be via procedures and automation of tasks. The real practices include for instance "workarounds" for problems of the official procedures, and may contribute to resilience and/or damage it, by creating gaps in the defences planned by designers and managers. It is appropriate to consider differences identified by "resilience engineering" authors between the "resilience engineering" and the older "high reliability organisation" movement. Perhaps the most cited paper [783] from the latter discussed how flight operations on U.S. Navy aircraft carriers achieved high success rates with remarkably good safety. This paper focused on four factors: "self-design and self-replication" (processes are created by the people involved, in a continuous and flexible learning process), the "paradox of high turnover" (turnover of staff requires continuous training and conservatism in procedures—both seen as generally positive influences—but also supports diffusion of useful innovation), "authority overlays" (distributed authority allowing local decisions by low-ranking people as well as producing higher level decisions through co-operation and negotiation), "redundancy" (in the machinery and supplies but also in overlapping responsibilities for monitoring and in built-in extra staffing with adaptability of people to take on different jobs as required). In contrast, a paper about how "resilience engineering"

[680] differs from this approach refers to healthcare organisations and how their culture and lack of budgetary margins severely limit the applicability of the four factors claimed to be so important on aircraft carriers; it points at the potential for improving resilience by, for instance, IT systems that improve communication within the organisation and thus distributed situational awareness and ability to react to disturbances. Another valuable discussion paper [586] emphasises the steps that lead from the general sociological appreciation of common issues—exemplified by the "high reliability organisation" literature—to an engineering approach with considerations of cost and effectiveness in detail.

1.4 Resilience and Fault Tolerance Against the Unexpected

We see that a frequently used argument for both fault tolerance (or "resilience", seen as going beyond standard practices of fault tolerance in a given community) in technical systems and more general "resilience" in socio-technical systems is based on these being broad-spectrum defences. Given uncertainty about what faults a system may contain or what external shocks and attacks it has to deal with, it seems better to invest in flexible, broad-spectrum defensive mechanisms to react to undesired situations during operation, rather than in pre-operation measures (stronger components, more design verification) that are necessarily limited by the designers' incomplete view of possible future scenarios.

1.4.1 Competing Risks; the Risk of Complex Defences

This argument can, however, be misleading. It is true that general-purpose redundancy and/or increased resources (or attention) dedicated to coping with disturbances as they arise, or to predicting them, can often deal with threats that designers had not included in their scenarios. But there will also be threats that bypass these more flexible defences, or that are created by them. An example can be found in the evolution of modular redundancy at the level of whole computers. The "software implemented fault tolerance" (SIFT) concept in the 1970s [384], the precursor of many current fault-tolerant solutions, responded to the fact that one could affordably replicate entire computations running on separate computers, so that the resulting system would tolerate any failure of any hardware or software component within a single computer (or communication channel). This was certainly a more general approach than either more expenditure on fault avoidance without redundancy, or ad-hoc fault tolerance for foreseen failures of each component in a single computer. It was a more powerful approach in that it may well tolerate the effects of more faults, e.g., some design faults in the assembly of the computer or in its software (thanks to loose synchronisation between the redundant computers [390]). But the SIFT approach also ran into the surprise of "inconsistent failures": the same loose, redundant organisation that gives

the system some of its added resilience makes it vulnerable to a specific failure mode. A faulty unit, by transmitting inconsistent messages to other units, could prevent the healthy majority of the system from enforcing correct system behaviour. To tolerate a single computer failure might require four-fold redundancy (and a design that took into account this newly discovered problem) rather than three-fold as previously believed. This was an unexpected possibility, although now, with experience grown from its discovery, it is easy to demonstrate it, using a simple model of how such a system could operate.

Other events that may surprise designers may be unexpected hardware failure modes; operators performing specific sequences of actions that trigger subtle design faults; new modes of attack that "create" new categories of security vulnerabilities; threats that bypass the elaborate defences created by design (ultra-high availability systems go down because maintenance staff leave them running on backup batteries until they run out, testing at a nuclear power plant involves overriding safety systems until it is too late for avoiding an accident (the Chernobyl disaster), attackers circumvent technical security mechanisms in ICT via social engineering); in short, anything that comes from outside the necessarily limiting model of the world that the designers use. Some such surprises arise from incomplete analysis of the possible behaviours of a complex system and its environment (cf the Ariane V first-flight accident [592]). Perhaps the incompleteness of analysis is inevitable given complexity, and indeed there is a now common claim that accidents—at least in "mature" organisations and engineered systems—tend to originate from subtle combinations of circumstances rather than direct propagation from a single component failure [722]).[2] On the other hand, designers also *choose* "surprises" to which their systems will be vulnerable: they explicitly design fault tolerance that will not cope with those events that they consider unlikely, trading off savings in cost or complexity against increases in risk that are (to their knowledge) acceptable.

In the ICT area, it is tempting to see "surprises" as manifestations of designer incompetence, and indeed, in a rapidly evolving field with rapidly increasing markets, many will be ignorant about what for others is basic competence. But there is also a component of inevitable surprises. In other areas of engineering it has been observed that the limits of accepted models and practices are found via failure [725, 921], usually of modest importance (prototype or component tests showing deviations from model predictions, unexpected maintenance requirements in operation, etc), but sometimes spectacular and catastrophic (the popular textbook examples—the Tacoma Narrows bridge, the De Havilland Comet).

[2] Although many authors point out that accidents caused by single component failures are still common. A component failure occurs in a system design that happens to omit those defences that would prevent that specific failure from causing an accident.

1.4.2 Quantifying Surprises?

Thus, the argument that a more "resilient" design—more open-ended forms of redundancy—offers extra protection is correct, but when it comes to estimating *how much* extra protection, or *which form* of redundancy will be more effective—when we need measurement and quantitative assessment—there is a difference between threats. There is a range of degrees to which quantitative reasoning is useful, perhaps best illustrated via examples. For a well known and frequent hardware failure mode, we may be able to trust predictions of its frequency, and thus predict the system reliability gain afforded by a specific redundant design, if some other modelling assumptions are correct. For other forms of failure, we may have very imprecise ideas about their frequency—for instance, this usually applies, at the current state of practice, to software failures in highly reliable systems—and yet, we can decide which designs will tolerate specific failure patterns, and via probabilistic modelling even decide whether a design is more resilient than another one given certain plausible assumptions. Last, there are surprises that violate our modelling assumptions. Designers can try to reduce them by keeping an open mind, and making the system itself "keep an open mind", but have no indication of how successful they are going to be. In the case of organisations, it may well be, for instance, that organisational choices that improve resilience against certain disturbances will be ineffective or counterproductive against others [933].

Insofar as resilience is obtained by making available extra resources, limits on resources demand that designers choose against which threats they will deploy more redundant resources. Limits on resources also recommend more flexible designs, in which these resources can deal with more different challenges. Again, these qualitative considerations demand, to be applicable to concrete decisions, quantification (at least adequate to support rough comparisons) of the risk and costs of different solutions.

This set of considerations has highlighted many areas where measurement and assessment of resilience or fault tolerance are desirable, and started to evoke a picture of measures that may be useful and of the difficulties they may involve. The discussion that follows looks at choices of attributes to measure, and difficulties of measurement and prediction, in some more detail, taking a viewpoint inspired by "hard" quantification approaches in engineering and considering some of the issues created by extension towards more complex socio-technical systems.

1.5 Quantifying Resilience: Its Attributes, and Their Possible Measures

In quantitative assessment there are always two kinds of potential difficulties: defining measures that usefully characterise the phenomena of interest; and assessing the values (past or future) of these measures.

About the first difficulty, dependability and resilience are broad concepts encompassing multiple attributes, so that there are multiple possible measures. The discussion that follows will take for granted that there are many dependability attributes of potential interest, which are different and may well be in conflict under the specific constraints of a certain system: for instance, pursuing safety—ability to avoid specific categories of mishaps—may conflict with the pursuit of availability—the ability to deliver service for a high fraction of the time (see for instance [63] for a high-level set of definitions). Irrespective of the specific dependability attribute of interest, we will summarily characterise categories of measures related to fault tolerance and resilience, with some discussion of their uses and difficulties in measurement and prediction.

The categories will be introduced in terms of "systems" (meaning anything from a small gadget to a complex organisation) that have to behave properly despite "disturbances" (an intentionally generic term, to cover component faults inside the system, shocks from outside, overloads, anomalous states, no matter how reached).

The sections that follow

- first discuss categories of measures in common use in quantitative reasoning about ICT, both as measures at whole-system level and as parameters, describing components and their roles, in mathematical models for deriving such whole-systems measures:
  - measures of dependability in the presence of disturbances, which may be estimated empirically in operation or in a laboratory, or through probabilistic models (as functions of measures at component level), as discussed in other chapters of this book
  - measures of the amount of disturbances that a system can tolerate, typically obtained from analysing a system's design
  - measures of probability of correct service given that a disturbance occurred ("coverage factors"), typically estimated empirically, often in a laboratory
- and then proceed to examine more speculative areas:
  - proposed predictors of resilience in socio-technical systems
  - more detailed measures that discriminate between different forms of "resilient" behaviours.

While pointing out differences between categories of systems and types of "resilience", the discussion will identify problems that they share and that may recommend importing insights from some areas of study to others.

1.5.1 Measures of Dependable Service Despite Disturbances

The first category of measures that give information about resilience are simply measures of dependability of the service delivered by a system that is subject to disturbances. The better the system worked despite them, the more resilient it was.

Indeed, a question is why we would want to measure "resilience" or "fault tolerance" attributes, rather than "dependability" attributes. The former are just means for achieving the latter.

For instance, an availability measure for a function of a system, obtained over a long enough period of use in a certain environment (pattern of usage, physical stresses, misuse, attacks etc), will be a realistic assessment of how well that function tolerates, or "is resilient" to, that set of stresses and shocks.[3]

This kind of measure is certainly useful when applied to documenting past dependability. It will be useful, for instance, in invoking a penalty clause in a contract, if the achieved availability falls short of the level promised. It will also have some uses in prediction. Suppose that the system is a computer workstation used for well-defined tasks in a relatively unchanging environment. A robust measure of past availability ("robust" may imply for instance repeating the measure over multiple workstations of the same type, to avoid bias from variation between individual instances) will be trusted to be a reasonable prediction of future availability (if the environment does not change). Measures on two types of workstations will be trusted to indicate whether one will offer substantially better availability than the other.

The Difficulty of Extrapolation

If we wish to compare systems (workstations, in this example), that have not been operated in the same environment, we will sometimes define a reference load (of usage as well as stresses etc)—a "benchmark" workload and stress (or *fault*) load, in the current IT parlance (see Chap. 14). Here, the broader "resilience" literature about engineering and socio-technical systems has to confront difficulties that are also evident for strict computer dependability evaluation [616], but with differences of degree. These difficulties can be generally characterised as *limits to the extrapolation of measures* to environments that are different from those where the measures were obtained. If a system copes well in the presence of one type of disturbances but

[3] A conceptual problem arises here, which will recur in different guises throughout this discussion. To use an example, suppose that two computers are made to operate in an environment with high levels of electromagnetic noise. Of the two, computer A is heavily shielded and mostly immune to the noise. The other one, computer B, is not, and suffers frequent transient failures, but always recovers from them so that correct service is maintained. The two thus prove equally dependable under this amount of stress, but many would say that only B is so dependable *thanks to* its "resilience": A just avoids disturbances; only B "bounces back" from them. Should we prefer B over A? Suppose that over repeated tests, B sometimes fails unrecoverably, but A does not. Clearly, A's lack of "resilience" is then not a handicap. Why then should we focus on assessing "resilience", rather than dependability? Or at least, should we not define the quality of interest (whether we call it "resilience" or not) in terms of "correct behaviour despite pressure to behave incorrectly"? An answer might be that the resilience mechanisms that B has demonstrated to have will probably help it in situations in which A's single-minded defence (heavy shielding) will not help. But then the choice between A and B becomes an issue of analysing how much better than A B would fare in various situations, and how likely each situation is. Measures of "resilience" in terms of recovery after faltering are just useful information towards estimating measures of such "dependability in a range of different situations".

less well with another type, changing the relative weights of these two types of disturbances will change the degree of dependability that will be observed. There will not even be a single indicator of "stressfulness" of an environment, so that we can say that if a system exhibited—say—99% availability under the benchmark stress, it will exhibit *at least* 99% availability in any "less stressful" environment [739]. Likewise, we won't be able to trust that if system A is more dependable (from the viewpoint of interest: e.g., more reliable) than system B in the benchmark environment, it will still be more dependable in another environment. An extreme, but not unusual case of the extrapolation problem is the difficulty of predictions about systems that are "one of a kind" (from a specific configuration of a computer system, to a specific ship manned by a specific crew, to a specific spontaneous, temporary alliance of computers collaborating on a specific task in the "future internet") or will be exposed to "one of a kind" situations: that is (to give a pragmatic definition), systems or situations for which we have no confidence that the measures taken elsewhere, or at a previous time, will still prove accurate. Again, extreme examples are easily found for the human component of systems: an organisation that appears unchanged, after some time from a previous observation, in terms of staff roles, machinery, procedures, may in reality have changed heavily due to staff turnover, or ageing, or even just the experience accumulated in the meantime (for instance, a period without accidents might reduce alertness). Here arises the first reason for going beyond whole-system dependability measures: they do not produce an understanding of *why* a system exhibits a certain level of dependability in a given environment—how each part of the system succumbed or survived the disturbances, which behaviours of which parts accomplished recovery, why they were effective—which could turn into a model for predicting dependability as a function of the demands and stresses in other environments.

Another problem with extrapolation is often created intentionally, as a necessary compromise. If we want a benchmark to exercise the whole set of defences a system has, we need the environment to "attack" these defences. This may require the benchmark load to condense in a short time many more stress events than are to be expected in real use; but some aspects of resilience are affected by the frequency of stresses. If the system being "benchmarked" includes people, their alertness and fatigue levels are affected. If it involves slow recovery processes (say, background processes that check and correct large bodies of data), an unrealistically high frequency of disturbances may defeat these mechanisms, although they would work without problems in most realistic environments.

Last, there is the problem of resilience against *endogenous stresses*. These exist in all kinds of systems: a computer may enter an erroneous state due to a software design fault being activated or an operator entering inappropriate commands; a factory may suffer from a worker fainting, or from a fire in a certain piece of machinery; and so on. If we wish a common benchmark to measure resilience against these kinds of disturbance, it will need to include some simulation of such events. But this may produce unfair, misleading measures. Perhaps a computer that has very little tolerance to errors caused by internal design faults has been designed this way for the right reasons, since it has no design faults of the types that it cannot tolerate; the less a

computer interface tends to *cause* operator errors, the less the computer needs to tolerate them; the less a factory tends to cause workers to become ill on the job, the less it needs to operate smoothly through such events; etc.

This unfairness also has a beneficial aspect, though: it allows a benchmark to give at least some information about resilience against the unexpected or unplanned-for disturbances. The benchmark deals with hypothetical situations. What if in a factory where nobody ever becomes ill, one day somebody does? What if the computer does have unsuspected design flaws? Likewise, modern regulations require many safety measures for all systems of a certain kind, irrespective of the probability, for a specific system, of the situations in which they would be useful. In these circumstances, a dependability or safety "benchmark" (from a fault injection experiment in a computer to an emergency drill in a factory) verifies that certain precautions are in place, and thus certain stresses are likely be tolerated if they were ever to happen. However, engineering for better dependability under a benchmark situation does not necessarily improve dependability in any operational situation different from the benchmark.

1.5.2 Measures of Tolerable Disturbances

A type of attributes that often allow simple and intuitive measures, and thus are heavily used, is the extent of deviation (or damage or disturbance) that a system can tolerate while still later returning to the desired behaviour or state (or still preserving some invariant property about its behaviour, e.g., some safety property: choosing different invariants will define different measures).

Thus, in ICT it is common to characterise a certain fault-tolerant computer design as able to mask[4] (without repair) up to k faulty components; or a communication code as able to detect (or to reconstruct the original message despite) up to t single-bit errors; or that a user interface will tolerate up to m erroneous inputs in one transaction; etc. Likewise, in the world of larger systems, we can rate a ship as being able to self-right from a tilt of so many degrees from the upright position; or a factory's staffing level as being calculated to allow for so many absences without loss of productivity. In ecology, a proposed measure of "resilience" of an environment is the size of a basin of attraction, in its state space: the distance by which the environment's state may be moved from a stable point without becoming unstable and moving into another basin of attraction (this distance measure is proposed to be used with a complementary measure of "resistance": the "force" needed for a given shift in the state space) [174].

To generalise, this set of attributes, and their measures, are about how far the object of interest can be pushed without losing its ability to rebound or recover; or how quickly it will rebound, or how closely its state after rebounding will resemble the state before the disturbance. To reason properly about these attributes of a system, it is important to recognise them as separate: system A may be "more resilient" than

[4] "Masking" usually meaning that the externally observed behaviour of the system shows no effect of the fault.

system B from one of these viewpoints, and "less resilient" from another one; for instance, A may be slower than B in recovering from a disturbance of a certain size, but able to recover from a more extreme disturbance than B can.

A great advantage of this type of measures is that for many ICT systems they are easy to obtain directly from their designs: so long as the implementation matches the design in some essential characteristics, we know that certain patterns of faults or disturbances are tolerated. These measures are also typically robust to the extrapolation problem.

If "measuring" on the design is unsatisfactory (for instance we expect the implementation to have flaws; or the required measure is too complex to calculate), we would rely on observations of the system in operation. There may be difficulties in obtaining enough observations of "disturbances" close to the limit, in knowing where the limit is (for systems that should not be tested to destruction), and in deciding whether the system's resilient reaction is deterministic, that is, whether observing successful recovery from a certain extent of disturbances allows us to infer 100 % probability of recovery. Again, socio-technical systems offer the most striking examples of the doubts that can affect estimates of these measures.

A limitation of these "maximum tolerable disturbance" measures, even for systems where they are easy to obtain, is that we may well be interested in characterising how well a system rebounds from *smaller* disturbances. For instance, given a form of fault tolerance that allows for some degradation of service, we may then want to measure not just how far the system can be pushed before failing altogether, but the relationship between the size of disturbances and the degradation of performance. For instance, for a network (of any kind) one might measure the residual throughput (or other measure of performance) as a function of the amount of network components lost (or other measure of faults or disturbances); this kind of function has been proposed [365] for resilience of critical infrastructures, leaving open the question of which single-number characterisation (if any) of these curves would be useful in practice. We will return later to characterisations of resilience as a function rather than a single, synthetic measure.

1.5.3 Measures of "Coverage Factors"

If we recognise that for most systems of interest the resilient behaviour is non-deterministic in practice,[5] we are no longer interested in *whether* the system will rebound from a disturbance but in the *probability* of it successfully rebounding; or perhaps the distribution of the time needed for it to return to a desired state; or other probabilistic measures. Thus in fault-tolerant computing we talk about the

[5] That is, even for many deterministic systems, their behaviour is complex enough that the knowledge we can build about them is only statistical or probabilistic. For instance, many software systems (deterministic in intention) have a large enough state space that many failures observed in operation appear non-deterministic—they cannot be reproduced by replicating the parts of the failure-triggering state and inputs that are observable [390].

"coverage" factor of a fault-tolerant mechanism, defined as the probability of the mechanism successfully performing its function in response to a disturbance (e.g., detecting a data error, or recovering from it), conditional on the disturbance (e.g., the data error) occurring; or we talk about the probability distribution of the latency of a component fault (i.e. of the time needed to detect it) rather than of a single numerical estimate.

Coverage factors are especially attractive as true measures of resilience. For instance, if we estimate a probability of a disturbance being tolerated so as not to cause system failure (a coverage factor, c), and know the frequency f of disturbances, then, in a simple scenario with rare disturbances, $(1 - c) * f$ would give us the frequency of system failures. The frequency of system failures is a measure of dependability (reliability), and one can see, for instance, that to improve reliability in this scenario one needs either to reduce the frequency of disturbances or to increase the coverage factor: the latter does indeed represent resilience, how well the system responds to adversity. And, as a concrete advantage, this relationship between a coverage factor and failure frequency seems to support extrapolation of dependability assessment to a different environment: I could estimate the former in the laboratory, usually with artificially frequent disturbances, to make measurement easier, and then I could extrapolate to any environment where the same disturbances occur more or less frequently. This is the basis for the predictive use of fault injection as described in Chap. 13 or dependability benchmarking as described in Chap. 14.

Even with complex systems in which multiple components and mechanisms co-operate to achieve resilience, probabilistic models, as described elsewhere in this volume, allow predictions of the probability of successfully resilient behaviour and hence of dependability measures as functions of coverage factors and of frequency of disturbances (internal faults or externally generated shocks).

However, this possibility of extrapolation is actually severely limited. Importantly, the probability of tolerating a disturbance will be a function of the type disturbance that occurred. So, all "coverage" measures have to be defined with respect to some stated type, or mix, of faults or disturbances; and the difficulties of extrapolation that characterised measures of dependability under stress also affect, in principle, measures of coverage. In particular, the desirability but also the limits of "benchmark" scenarios apply when estimating coverage factors just as when measuring a dependability measure [739].

1.5.4 Measures of Socio-Technical Resilience

Since we are comparing the understanding of resilience with respect to different categories of systems, and the categorisation above is derived from examples at the simple end of the spectrum, it is useful to compare with proposed measures in the areas of complex socio-technical systems. We take as an example the list of attributes of resilience in socio-technical systems proposed by Woods [941]; we can relate them

to the categories given above, as well as consider how amenable they are to precisely defined measures. These attributes are:

- "buffering capacity", which is essentially an "extent of tolerable disturbances" as discussed above. The potential difficulties only concern how easily this can be captured in practically usable measures;
- "flexibility versus stiffness: the system's ability to restructure itself in response to external changes or pressures". It is not clear how this could be measured. For instance, to measure flexibility in the observed operation of a system, we would need to decide which forms of "restructuring" were actually useful, without the benefit of checking how the crisis would develop if the restructuring had not taken place. So, the literature tends to describe this form of "flexibility" through scenarios or anecdotes;
- "margin: how closely or how precarious the system is currently operating relative to one or another kind of performance boundary", again related to "extent of tolerable disturbances". This has often useful definitions in technical systems, for instance we can define an acceptable maximum load on a network before it goes into congestion, or the minimum required set of functioning components necessary for basic services, while in socio-technical systems it is often difficult to identify what terms like "stretched to breaking point" may mean, and what measures of "distance" from this point may be appropriate;
- "tolerance: how a system behaves near a boundary—whether the system gracefully degrades as stress/pressure increase or collapses quickly when pressure exceeds adaptive capacity". This has parallels in many technical areas, and certainly in ICT, where "graceful degradation" is a frequent requirement, but for which no textbook, standardised measure exists.

1.5.5 Measuring the Supposed Determinant Factors of Resilience

When trying to assess dependability (and resilience) in the face of threats that cannot be predicted in detail, a proposed approach relies on identifying factors that are believed to enhance resilience. When dealing with well-understood risks, this exercise may take the form of simple design analysis. In many cases, assessment can rely on the combination—via a probabilistic model—of analysing which defensive mechanisms are in place, estimates of their coverage factors, and estimates of the probability distributions of disturbances to which they will need to react. There are of course difficulties with all these estimates, which qualify the confidence one can have in predictions obtained this way. But when dealing with the human and social determinants of system behaviour, the conjectured determinant factors of resilience often have a "softer" or at least more complex character. The coverage and component reliability parameters of a model for a complex socio-technical system, and even the model itself, would be often too difficult to establish with any confidence. Only empirical observations of system resilience would then be trusted.

A concern in the "resilience engineering" literature is that one tends to judge organisations on their past performance, but these "measures of outcomes" may lack predictive power: success in the past is no guarantee of success in the future, due to the extreme extrapolation problems mentioned above. Hence a search for "leading indicators" that can be used to assess future resilience. Many of these are cited in the literature. For instance, a review [432] lists measures of "Management commitment, Just culture, Learning culture, Opacity, Awareness, Preparedness and Flexibility", of "Empowerment, Individual responsibility, Anonymous reporting, Individual feedback, […]" for individual workers and of "Organizational structure, Prioritizing for safety, Effective communication" for organisations (citing [388]); and others.

Such factors are commonly believed to be important in determining how well an organisation will perform from the safety and resilience viewpoints. So, informed judgements about how "resiliently" organisations will react to stresses will benefit from considering these "indicators"; if the indicators were reliable, an organisation might want to identify reasonable target values and levels of trade-offs among them. But objective measures of such attributes are difficult to define. Different systems can be ranked on ordinal scales with regard to attributes of interest, or specific numerical, objective measures can be used as proxy measures if shown empirically to correlate with desired behaviours. For instance, [389], studying the safety of ship operation, reports a massive effort in which factors believed to indicate "safety culture" were estimated by anonymous surveys of individuals; the research goal is to check how well these proxy measures correlate to observed safety performance (such as records of accidents, near misses, negative reports by competent authorities). If good correlation were found, some function of the "leading indicators" could be used for early warnings of accidents being too likely on a certain ship. On the other hand, predictive models akin to those described in this volume (e.g., Chap. 7), based on such measures and suitable for informing design of these systems, for instance answering questions like "To what extent should power be devolved to workers in this system so that the positive effects outweigh the negative effects?", appear unfeasible.

Precedents for emphasis on "determinant factors" of desired characteristics exist in all areas of engineering, as sets of mandated or recommended practices. A pertinent example is in standards for safety-critical computing, e.g., [461], where sets of good practices are recommended or required to be applied in developing and verifying software, as a function of the criticality of the software's functions. This is a reasonable approach, in principle, and yet checking that these practices were applied is a poor substitute for directly checking that the product has acceptably low probability of behaving unsafely: the former (good practice) does not imply the latter (safe enough behaviour). The difficulties are twofold: there is no clear knowledge of how much these practices, and their possible combinations, tend to help; and we should expect that (comparable) systems that are equal in the extent of application of these practices may still differ in the achieved results (the levels of dependability).

Indeed, many authors in the "resilience engineering" literature are wary of attempts at quantification, applied to complex systems, as liable to oversimplify the issues and divert management effort towards achieving required values of measures that have the "advantage" of concrete measurement procedures but no guaranteed

relationship to outcomes. Others have used quantitative modelling for illustration and general insight, borrowing physics-inspired formalisms for modelling complex systems at a macroscopic level [311].

1.5.6 More Detailed Characterisations of Resilience

Two important topics that have emerged in the discussion so far are: the difference between tolerance/resilience for "design base", expected disturbances and for unexpected or extraordinary (excluded by design assumption) ones; and the possible need to characterise not just the size of the tolerable stresses, but more detail about the resilient behaviour in response to different levels or patterns of stresses.

In this latter area, one can look for measures like *performability*, defined [646] as the set of probabilities of the "levels of accomplishment" of a system's function,[6] or functions like network throughput as a function of loss of components. These options are no sharp departure from dependability modelling approaches that are well established in ICT.

While these measures are meaningful, authors have been looking, as exemplified in the previous section, for ways to characterise "resilient" behaviour in a more detailed fashion, although accepting that the result may be qualitative insight rather than models suitable for prediction.

To discuss the various parameters that may characterise resilience in an organisation, Woods and Wreathall [942] use the "stress-strain" diagram used in material science, as in Fig. 1.1. With materials, the y axis represents the "stress" applied to a sample of the material (e.g., tensile force stretching a bar of metal), and the x axis represents the degree of stretch in the material ("strain"). When tested, the typical building material will exhibit a first region of linear response (the stretch is proportional to the force applied), followed by a less-than-linear region, and finally by quick yielding that leads to breaking. As it moves from the linear to the sub-linear region, the material also moves from elastic behaviour, where the original size will be regained when the stress is removed, to permanent deformation. A qualitative analogy with organisations is made, in terms of "a uniform region where the organization stretches smoothly and uniformly in response to an increase in demands; and an extra region (x-region) where sources of resilience are drawn on to compensate for non-uniform stretching (risks of gaps in the work) in response to increases in demands". Thus in the "extra region" it is assumed that an organisation that successfully self-modifies shifts onto a new curve, that departs from the now-decreasing main curve and gives some extra amount of increase in tolerated "stress" for extra "strain", so as to be able to tolerate stresses beyond its "normal" maximum.

So, these authors identify a region of "orderly" adaptation to increasing stress (in some cases one might identify measures of both stress and strain with an

[6] If "accomplishment" has a numerical measure, e.g., throughput of a system, the system's performability is defined by the probability distribution function of this measure.

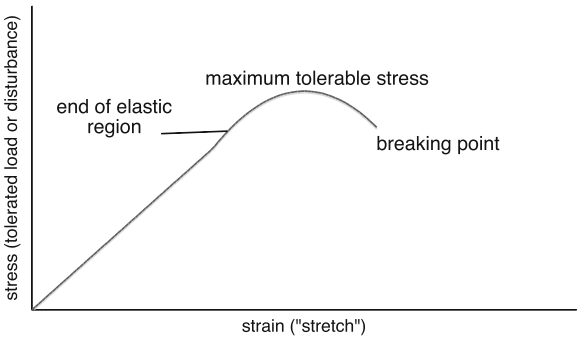

Fig. 1.1 Stress-strain diagram

approximately linear relationship, e.g., increased inflow of patients to a hospital being covered by increasing work hours within established procedures). Beyond this maximum, the cost-effectiveness of use of resources decreases and a maximum exists, beyond which extra stress can only be tolerated by some kind of reconfiguration of the organisation, e.g., mustering extra resources or freeing them by changes of operation mode.

This view suggests sets of attributes that can be measured to characterise the response of the system, like the size of the "uniform" range, and the extra stress that can be tolerated before the degeneration into failure. The above authors identify as especially important the ability of an organisation to manage smoothly transitions between regions, and its "calibration", defined as its ability to recognise in which region it is operating, so that reconfiguration is invoked when necessary (and presumably not too often: we note that in many real situations, the ability to assess how well calibrated they were for past decisions is limited. One cannot always tell whether a decision to restructure to avoid catastrophic failure was really necessary—especially in view of the uncertainty that the decision maker normally faces in predicting the future). They rightly claim that the stress-strain analogy for organisation behaviour is a first step in clarifying some of the attributes that characterise resilient behaviour (hence also a first step towards quantitative modelling) and importantly highlight the difference between "first-order" and "second-order" adaptive behaviour—the "normal stretching" of the organisation's design in the uniform region, versus the more radical restructuring to work beyond the "normal" limit—but note the limitation of representing "stress" as a unidimensional attribute, and the need for further work. A limitation that seems important is that this kind of graph implicitly assumes that the stress-strain relationship can be plotted as independent of time. This matches well those measurement processes for the strength of materials in which stress is increased slowly, moving between states of equilibrium at least up to the maximum of the curve. If the timing of the applied stimulus (as e.g., with sharp impact or repetitive stress) makes a difference in how the material reacts, additional properties can be studied, possibly requiring additional measures. In organisations (or for that matter in computers), many of the stresses may need to be characterised in terms of dynamic

characteristics, or need to be defined in practice in terms of timing characteristics of events.

Considering the time factor may also bring into play other aspects of self-stabilisation, and other necessary design trade-offs, which can be illustrated by analogy with other engineering examples, outside the science of materials. For instance, making a ship more "stable" (increasing its metacentric height, so that it will self-right more promptly after heeling to one side) makes it also more liable to roll at higher frequency following the tilt of the waves, a characteristic that can reduce the effectiveness of the crew, make a warship unable to use its weapons, etc. Likewise, all "resilience" that relies on detecting (or predicting) component failures or shocks must strike a compromise between the risk of being too "optimistic"—allowing the situation to deteriorate too far before reacting—or too "pessimistic"—reacting too promptly, so that false alarms, or reactions to disturbances that would resolve themselves without harm, become too much of a drain on performance or even damage resilience itself.

1.6 Conclusions

A theme running through this survey has been that as fault tolerance (or resilience), that is, dynamic defences, exist in all kinds of systems, the measures that may be appropriate for studying them also belong to similar categories and the difficulties in defining measures, in performing measurements, and in predicting the values of measures also belong to common categories. Interest in studying and/or in extending the use of fault tolerance or resilience has expanded of late in many areas,[7] and we can all benefit from looking at problems and solutions from different technical areas. I gave special attention to the "resilience engineering" area of study, since its choice of topic problems highlights extreme versions of measurement and prediction problems about the effectiveness of "resilience" that exist in the ICT area. In all these areas there are spectra of prediction problems from the probably easy to the intractable. The "resilience engineering" movement has raised important issues related to the measurement and prediction of "resilience" attributes. One is simply the recognition of the multi-dimensionality of "resilience". For instance, Westrum [933] writes: "Resilience is a family of related ideas, not a single thing. [...] A resilient organization under Situation I will not necessarily be resilient under Situation III [these situations are defined as having different degrees of predictability]. Similarly, because an organization is good at recovery, this does not mean that the organization is good at foresight".

The boundaries between strict technical ICT systems and socio-technical systems are fuzzy, and for many applications the recognition of social components in deter-

[7] U.S. Navy aircraft carriers exploited redundancy for safety long before Rochlin and his co-authors studied it. On the other hand, their study prompted more organisations to recognise forms of redundancy in their operation, and protect them during organisational changes, and/or to consider applying redundancy.

mining meaningful assessment of dependability is important [769]. Concerns about improving measurement and quantitative prediction are often driven by the concrete difficulties in applying existing methods in new systems: just as increasing levels of circuit integration and miniaturisation made it unfeasible to monitor circuit operation at a very detailed level via simple probes and oscilloscopes, so the deployment of services over large open networks and through dynamic composition may create new difficulties in measuring their dependability. More general problems may arise, however: do we need to choose appropriate new measures for characterising the qualities of real interest? If they are amenable to measurement in practice, to what extent will they support trustworthy predictions? To what extent may the benefit of "reasonably good" measures (perhaps acceptable proxies for the "truly important" ones) be offset by natural but undesirable reactions to their adoption: designers and organisations focusing on the false target of achieving "good" values of these measures, perhaps to the detriment of the actual goal of dependability and resilience?

These questions underlie all assessment of resilience and dependability, but more markedly so as the socio-technical systems studied become less "technical" and more "social". Authors in "resilience engineering" have identified research problems in better characterising, even at a qualitative, descriptive level, the mechanisms that affect resilience. Quantitative measurement may follow. Quantitative predictive models may or may not be feasible, using results from the abundant research in modelling—at various levels of detail—the dependability of complex infrastructure and ICT; quantitative approaches from mathematical physics [311] may also yield insight even without predictive power. Research challenges include both pushing the boundary of the decision problems that can be addressed by sound quantitative techniques, and finding clearer indicators for these boundaries. There are enough historical examples of quantitative predictions proving misleading, and perhaps misguided, but we often see these with the benefit of hindsight. Perhaps most important would be to define sound guidance for "graceful degradation" of quantitatively driven decision making when approaching these limits: more explicit guidance for exploiting the advantages of measurement and quantitative prediction "as far as they go" but avoiding potential collapse into unrealistic, "pure theory"—driven decision making.

Acknowledgments This work was supported in part by the "Assessing, Measuring, and Benchmarking Resilience" (AMBER) Co-ordination Action, funded by the European Framework Programme 7, FP7-216295. This article is adapted from Chap. 15 of the "State of the Art" report produced by AMBER, June 2009.