HAL
open science

Exploring the Dynamics of Resilient Performance

Robert Wears

To cite this version:

Robert Wears. Exploring the Dynamics of Resilient Performance. Business administration. École Nationale Supérieure des Mines de Paris, 2011. English. NNT : 2011ENMP0059 . pastel-00664145

HAL Id: pastel-00664145

https://pastel.hal.science/pastel-00664145v1

Submitted on 29 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n° 432 : Sciences et Métiers de l'Ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité " Science et Génie des Activités à Risques "

présentée et soutenue publiquement par

Robert L WEARS

le 27 octobre 2011

Exploring the Dynamics of Resilient Performance

Directeur de thèse : **Erik HOLLNAGEL**

Jury

M. Pierre FALZON, Professeur, CNAM	Rapporteur
Mme. Anne-Sophie NYSSEN, Professeur, Université de Liége	Rapporteur
M. Erik HOLLNAGEL, Professeur, CRC, Mines ParisTech	Directeur de Thése
M. Jean-Christophe LE COZE, PhD, INERIS	Examinateur
M. Henning ANDERSEN, Professeur, DTU	Examinateur

T H È S E

MINES ParisTech
Centre de recherche sur les Risques et les Crises
Rue Claude Daunesse, BP 207 – 06904 Sophia Antipolis

Acknowledgements

This thesis is the result of a long chain of influences stretching back over many years; many of those I need to thank may have little idea of the role they played in my growth and development in the worlds of system safety and performance.

Over 10 years ago, Richard Cook introduced me to modern thinking on safety and performance in complex work. Through him I came to know the work of David Woods, Gary Klein, Ed Hutchins, Erik Hollnagel, Neville Moray, Sidney Dekker, Jens Rasmussen and a host of others. He has been unfailingly gentle but firm in his criticisms and generous with his time; without his initial direction and encouragement, I would not have discovered this (to me) new world.

My supervisor, Erik Hollnagel, has endured a lot in my fits and starts, patiently corrected misunderstandings and tempered naïve enthusiasms. His calm, erudite and reflective thinking set a standard for writing and mentoring I can only hope to approach.

Kathleen Sutcliffe introduced me to the field of research on high reliability organisations, and more generally to the non-Cartesian/Newtonian forms of scientific activity. While I have chosen an engineering, rather than an organisational behaviour, framing for my work, I appreciate the contributions from that sector because I know they come from a thoughtful and informative place.

My department chair, David Vukich, was instrumental in this work. He created a setting in which faculty were permitted to explore and if possible, thrive, and very specifically, did the enabling leg-work to allow my sabbatical year to take place (since it had never before been attempted in the University of Florida College of Medicine). That experience – particularly the opportunity to meet and talk with many whose work I had been reading – set me onto this path.

Charles Vincent has been a friend and counselor during this time, and helped make my sabbatical time a transformative experience by giving me leave to explore a world inhabited primarily by psychologists and engineers, only occasionally visited by clinicians. The sabbatical experience began my transformation from clinician to

safety scientist, which transformation this thesis should signal – I will not say complete, as I expect to continue learning and changing.

My wife Dianne has borne the greatest burden during this effort. She endured the dislocations of a year abroad, long periods of my disappearance into reading and writing, and countless dinner-table and late evening expositions of my ill-formed but developing ideas. Without her unconditional support, nothing useful could have been accomplished. My family gave the same ungrudging support, particularly Bill and Ted, who helped repair my poor French translations.

I owe a great debt to my colleagues in the emergency department: physicians, nurses, and technicians. They have generously contributed their time and their stories, tolerated my strange questions and my love for taking pictures of the madness of the ER, and re-infused me with a sense of professionalism and a pride in craftsmanship that is shared by those who commit themselves to this difficult and sometimes dangerous work and do it well. I admire their ability to enact resilience in their work; if I were sick or injured, I would rather be their 10th patient than someone else's only patient.

Finally, I have benefited from numerous informative discussions, readings, and friendships that have grown out of the series of Resilience Engineering conferences. I recall these meetings being described as "short papers punctuated by long discussions", which to my mind is the best format for a scientific meeting. I look forward to many similar events in the future.

Notice

The opinions, findings, and conclusions or recommendations expressed in this document are those of the author and do not necessarily represent the official position or policies of Mines ParisTech or of the University of Florida.

Avertissement

Les opinions, constatations, conclusions ou recommandations exprimées dans ce document sont celles de l'auteur et ne reflètent pas nécessairement la position ou les politiques officielles de Mines ParisTech ou de Université de Floride.

Table of Contents

Acknowledgements ii

Notice iv

Avertissement iv

Table of Contents v

List of Figures viii

Foreword x

Chapter 1. Introduction 1

1.1 Motivating example 1

1.2 Advantages / goals of a model 3

1.3 Notes on models 7

1.4 A caution against reification 8

1.5 Overview 10

French summary of Chapter 1 12

Chapter 2. Models of resilience 15

2.1 Background 15

2.2 Synopsis 16

2.3 State space model 18

2.4 Stress-strain model 21

2.5 Parallels to other domains 22

2.5.1 California electric power transmission grid 23

2.5.2 IT system crash – situational on fundamental surprise 26

2.5.3 Rule-violating use of a tightly controlled drug 28

French summary of Chapter 2..... 31

Chapter 3. Abstraction 33

3.1 Classes of adaptations 33

3.2 Observed adaptations 35

3.2.1 Adaptations of exploitation 36

3.2.2 Adaptations of exploration 38

3.3 Stopping the system – a special case of exploration 39

French summary of Chapter 3..... 42

Chapter 4. Dynamics 45

French summary of Chapter 4..... 47

Chapter 5. Building a model..... 49

5.1 Work activity 50

5.1.1 Structure..... 50

5.1.2 Behaviour 53

5.2 Exploitation..... 57

5.2.1 Structure..... 58

5.2.2 Behaviour 60

French summary of Chapter 5..... 64

Chapter 6. Exploration 67

6.1 Fortuitous stopping 67

6.2 Diurnal variation as fortuitous stopping..... 72

6.3 Purposeful stopping 78

French summary of Chapter 6 82

Chapter 7. Inference and conclusion 84

7.1 Findings 84

7.1.1 Nonlinearity 84

7.1.2 Falling behind 85

7.1.3 Path dependence 85

7.1.4 Threshold phenomena 85

7.2 Implications and application to ‘free fall’ 87

7.2.1 A possible origin of ‘free fall’ 87

7.2.2 A possible means of recovery from ‘free fall’ 88

7.2.3 A possible strategy to reduce the risk of ‘free fall’ 89

7.3 What was gained by modeling 91

7.4 Limitations 91

7.4.1 Simplifications 92

7.4.2 System dynamics assumptions 94

7.5 Future Directions 95

French summary of Chapter 7 97

References 101

Appendix 1 Methods 112

Appendix 2 Model equations 113

Appendix 3 Related publications 120

List of Figures

Figure 1. Resilience state space diagram (modified from Hollnagel & Sundström, 2006). 20

Figure 2. Stress-strain representation of resilience (from Woods *et al*, 2006). 22

Figure 3. ED status board showing "board within a board" with reduced level of detail.....37

Figure 4. The basic work system. 51

Figure 5. Response of perfectly resilient work system to pulse overload 54

Figure 6. Response of a perfectly reslient system to increasing pulse challenges.....55

Figure 7. Response with overload degradation.....56

Figure 8. Brittle response to increasing overload. 57

Figure 9. Stress-strain plots of decompensation. 57

Figure 10. Exploitation of additional resources to compensate for overload. 58

Figure 11. The overload loop; a positive feedback loop.....59

Figure 12. The consumption loop; a negative feedback loop. 60

Figure 13. The restoration loop; a positive feedback loop. 61

Figure 14. Response of Margin to a pulse challenge at hour 60.....61

Figure 15. Adaptive resources mitigate overload shocks. 62

Figure 16. A sufficiently large and timely fortuitous decrement allows system recovery. 68

Figure 17. The ability to recover from a fortuitous decrement in demand critically depends on the magnitude of the decrease.....69

Figure 18. Effect of small changes in fortuitous decrease in demand on *Margin*. 70

Figure 19. Relationship between magnitude, duration, and timing of the smallest decrement in demand able to trigger recovery.....72

Figure 20. Diurnal variation in demand and workload.....73

Figure 21. A small pulse challenge at hour 60 (noon) creates serious disruption but the system can ultimately recover.....74

Figure 22. System load following a sub-threshold pulse challenge compared to normal operating conditions. 75

Figure 23. Response to subthreshold pulse challenge at hour 60. 76

Figure 24. Effect on *Margin* of subthreshold pulse challenge at hour 60. 76

Figure 25. Effect of partial stopping to restore *Margin*.....80

Foreword

The reader might reasonably wonder why an American would choose to obtain his graduate education, and to prepare and defend his thesis at a French institution of higher learning. In brief, this came about through a combination of intellectual appeal and fortunate circumstances.

Roughly ten years ago, relatively early in my study of the problem of safety in healthcare, I became aware of several (to me) important themes. First, it seemed to me that most US physicians were quite convinced that they both understood safety issues and had the theories, skills, and knowledge to address them effectively; I did not share in this high opinion. Second, in my readings I had become aware of alternative approaches to safety in complex socio-technical systems (such as healthcare) that strongly contrasted with the 'measure and manage' approach that seemed to be taken for granted in North America. For lack of a better term, I began to identify this school of thought to myself as 'European', which was serendipitously accurate, as the writers to whose work I was drawn came from the intellectual provenance of Europe, especially the French and Belgian 'work ecology' researchers of the late 20th century. I was strongly drawn to this thinking as more original, more insightful, more theoretically strong, and in the long run more productive than what was being accepted without question as the conventional wisdom about healthcare safety in the United States. (In particular, the distinction between 'task' and 'activity' – work as imagined *vs* work as performed – struck me as fundamentally important to safety, but was routinely ignored by mainstream thinking in American healthcare). Even the American researchers who had the greatest influence on me (*eg*, Richard Cook, David Woods, John Flach) were constantly referring me to works by Jens Rasmussen, Erik Hollnagel, Sidney Dekker, and others, and in fact would themselves be considered adherents of this 'European' school of thinking.

In the midst of this period of intellectual ferment and growth, I was fortunate enough to win the Society for Academic Emergency Medicine's *Scholarly Sabbatical Award*, with a proposal to spend a year across the Atlantic learning a distinctly different approach to safety and human performance from researchers and scholars on the

opposite side of the ocean. I made a base of operations in London for practical reasons, and spent that year reading, thinking, meeting and talking with those in Europe whose works I had read and learnt from, steeping myself in this school of thought, to which I had become a convert. This was a transformational experience; I discovered that not only was I drawn to Europe intellectually, but personally and socially as well, as I felt more at home in England, Belgium, or France than I did in the US (an impression which has continued until this day).

Finally, after my reluctant return to the US, the series of events occurred that I have described as "free fall" and that serve as the motivating example for this work. In trying to make sense of these events and the responses to them on multiple levels, I began writing about them, and in that activity found the resilience engineering framework to be the most expressive and useful approach. So, during the 2nd International Resilience Engineering Conference in Juan-les-Pines, Richard Cook suggested I approach Professor Hollnagel with the crazy idea that I should study with him, using resilience engineering as a framework for apprehending issues of safety and performance in healthcare and other complex endeavours. This I did, and he was kind enough to accept me; although he may have since had some regrets about it, I never have.

Chapter 1. Introduction

1.1 Motivating example

At around 1900 on 14 December 2005, the emergency department (ED) at Shands Jacksonville Health Center, a major, urban hospital and Level 1 Trauma Center in northeastern Florida, failed utterly and came to a complete halt. This was an extraordinary event; EDs are expected to perform at high levels 24 x 7 x 365, without any halts, breaks, or down-time, and are generally successful in meeting this expectation. They are staffed by a largely self-selected group of health professionals who take great pride in their ability to "roll with the punches", to keep on working effectively no matter what the circumstances. (This is exemplified by the title of a recent history of emergency medicine as a specialty – *Anyone, Anything, Anytime* (Zink, 2006)). In addition, the staff at this ED had long and frequent experience in responding effectively to high volumes of critically ill or injured patients, and in dealing with externally provoked medical and public health crises, such as hurricanes, wild fires, influenza epidemics, *etc*. But, on this evening, without any external trigger, the ED lost its ability to function. To use an everyday analogy from personal computers, the ED had crashed ("locked up" might be more accurate) and had to be re-booted.

What is difficult to express in prose is the profound sense of defeat and failure this event produced in the staff involved. As the lead attending physician on duty on this occasion, I was led to question my own capabilities in my chosen specialty, and to question our capabilities as an organization[1]. It challenged the very soul of our professional identity, because nothing remotely close to this sort of abject failure had ever occurred in my 32 years as a physician or in my 20 years of experience in this ED. Although I have never experienced combat, the analogy that most accurately expressed this situation seemed to be that of being overrun by the enemy in ground combat, where the only possible course of action was to hunker down in a hole until

[1] Overload is commonly viewed in healthcare as a phenomenon of novices, or advanced beginners, not experts; thus, the subjective experience of overload was personally challenging to the senior clinicians involved.

the worst was over, and then to gradually emerge and reorganize to begin operations again. This negative affect was reinforced by the strong feeling that the causes of the incident were internal to the medical center; that the ED had been put in an untenable situation not by external, uncontrollable events, but rather by the actions of our own leadership[2]. Black humour is common in emergency medicine (Arthur L Kellermann, 2010), and in the informal discussions of this event within the ED staff, the situation began to be referred to as "free fall" – a loss of control leading to rapid degradation of performance that could only be resolved after "hitting bottom" and restarting the system.

In trying to make sense of this event, I began to write about it, especially since in discussing the episode in formal and informal interviews with involved staff, I discovered another similar event in our ED had occurred about a month earlier (14 November 2005). This suggested that something more fundamental was going on, something more than an idiosyncratic and infortuitous intersection of mischances. These musings were first therapeutic, describing and venting, but eventually they became reflective, particularly in the context of what I had learned in my sabbatical year, and was continuing to learn. As I discussed these events with my colleagues and mentors in the months immediately following December 2005 (especially with Richard Cook and David Woods), I began to use resilience engineering as an interpretive framework, and found that it enabled me to make sense of that which heretofore had made no sense. Thus, the "free fall" episode serves as a motivating example for this work – a stimulus to a more general intellectual effort. In essence, these motivating data are a "theory fragment" ripe for elaboration and theory development (Davis, Eisenhardt, & Bingham, 2007).

Resilience can be defined as "the intrinsic ability of a system to adjust its functioning prior to, during, or following changes or disturbances, so that it can sustain required operations under both expected and unexpected conditions" (Hollnagel, 2011). We

---

[2] In a subsequent after-action-review (conducted only at the staff's insistence), our chief executive likened the "free fall" situation to the Battle of the Bulge; the involved staff bristled at this suggestion, because in that setting, personnel were put at risk by enemy action. They suggested a better analogy was Khe Sanh, where personnel were put at risk by their own leaders' decisions.

often talk of resilience as if it were something a system has, which implicitly denies its dynamic properties. In this work, I treat resilience as something a system *does*, or more specifically, a set of ways in which it does that which it does, and work toward understanding its dynamics, of how resilient performance plays out over time. (Interestingly, this change in framing is accompanied by a parallel change in syntax – resilience considered initially as a noun, then as an adjective, and finally as an adverb).

Specifically, I extend current theoretical models of resilience (Hollnagel & Sundström, 2006; Woods, Wreathall, & Anders, 2006) by developing system dynamic simulation models (Sterman, 2000) to represent theoretical concepts and relationships about resilient performance and its dynamics. The fundamental idea underlying this approach is the notion that the resilient stability we see in complex systems results from a dynamic equilibrium – one which requires constant inputs from its operators to maintain, and which if left alone, will collapse. I use this approach to explore the emergent properties of a theory of resilience dynamics.

1.2 Advantages / goals of a model

Before entering into the detailed exposition of the project, it seems reasonable to address the question: Why model? What, exactly, does a model provide? By addressing this issue before a detailed explication of the empirical and published data that informed the modeling exercise, and before a description of the model itself and its results, I hope to provide the reader a clearer guide to what will follow.

It is important to note that for the purposes of this project, the development of a model represents the development of a hypothesis – in a strong sense, the model *is* the hypothesis. The model ultimately becomes a theory fragment; and the project itself the inductive construction of theory.

The intent behind this attempt to model the dynamics of resilience is to capture general properties that can be used to understand how specific systems will behave when they encounter challenges or their adaptive capacity is degraded (Woods, 2011). A useful model would serve by providing support for the understanding of resilience

in the ways listed below. Where applicable, I have noted which of the 4 cardinal resilience activities (anticipating, monitoring, responding, learning) benefits directly.

- A model provides explanatory power, especially in a dynamic sense. There are already useful models of resilience (Hollnagel & Sundström, 2006; Woods & Wreathall, 2008), but they have tended to be static in the sense that while they indicate that systems do change states over time, they do not explain why a system changes when – what provokes or inhibits change at a given time – or when a change is imminent. Essentially, a model allows us to encapsulate learning about system behaviours (especially as a result of modifications), in the same sense that one might say that the structural diagram of a chemical compound encapsulates what is known about how it reacts with other compounds. So, fundamentally, a model represents and facilitates *learning*.
- A model would help articulate what has been called the "adjacent possible" (Johnson, 2010; Kauffman, 1995, 2000), the set of states (whether improved or degraded) to which a system might be prone to transition. By making more explicit this "shadow future, hovering on the edges of the present state of things" (Johnson, 2010), a model would facilitate *anticipation*. Dynamic models do not attempt to map causes directly onto consequences, but rather focus on generating the set of possible consequences that stem from a given set of causes; they are aimed at elucidating the general principles that govern how outcomes occur in a domain, without having to postulate specific causal events (Hollnagel, 1993).
- A model would suggest potential indicators of the loss of "margin for maneuver" (Pariès, 2011; Stephens, 2010; Stephens, Woods, Branlat, & Wears, 2011). For example, it would be extremely useful to system operators (and indeed, to those who depend upon the system) to be able to recognize that a sudden change in state (*eg*, a dramatic reconfiguration) is impending. The fact that reconfigurations can be either beneficial (and thus should be promoted if overdue) or deleterious (and thus should be forestalled if possible) adds to the importance of being able to recognize that the system is on the cusp of a state change. While being late with an adaptive reconfiguration

might possibly only be costly in terms of efficiency (although it could certainly also affect safety or survivability), flipping into a maladaptive state is unambiguously bad – particularly since such changes often exhibit hysteresis, *ie*, they are persistent long after the provoking circumstances have returned to their previous states. Thus, a model would help with the question of what are good ways of noticing that things are going to turn bad before they become bad. Thus, a model facilitates both *anticipating* and *monitoring*.

- An important advantage of a dynamic model is that it allows us to investigate the timing of interventions or adaptations. In many systems, timing is everything; premature interventions may be costly and harmful, while late interventions may be both costly and ineffective. So a model that could help operators know whether they should continue "just a bit longer" before changing, or should instead change right now would be useful in avoiding the chagrin, regret, and potential damage of realizing that they have already fallen behind the tempo of operations (Woods & Branlat, 2011a) and should have changed earlier. In addition, the knowledge that a change in state was impending, or even possible in a certain context, would allow operators to prepare for it. A model would be particularly valuable in understanding these dynamics, because managing timing effectively is exercised largely through intuition – *praxis*, the often inarticulate knowledge of knowing what best to do in a situation – and *praxis* is learnt largely through experience. In complex, expensive, hazardous systems subject to "wicked problems" (Rittel & Webber, 1973), trial and error learning from experience may not be possible – the first error may be the last trial. And often, nature is merciful, and thus stingy with experience about difficult problems – catastrophes, crises, and the like occur seldom if at all in well-designed and well-regulated systems. Interaction with a dynamic model would provide an opportunity to practice the timing of adaptations, or to identify guides in the sense of affordances for action. Thus, a model would facilitate *learning* and *responding*.
- Similarly, a model provides the capability to test possible responses and strategies, in affording a means of exploring adaptive strategies (and their

timings) in a safer, more peaceful manner by doing so in a model rather than in the real world.

- What is sorely needed in most hazardous systems is a better way for controlling the complex situations that may arise (Woods & Branlat, 2010), and a model is a step in that direction. It give operators at different levels of the system (Woods & Shattuck, 2000) a better way of understanding what is going on, a better way of guiding their attention to what is important, and a better way of predicting what is likely to happen as a result of particular adaptations.
- Ultimately, a model is an aide to learning. Many adaptations are initially adopted in desperation, and are essentially explorations of the unknown. Some of these, after operators have learned from their experience and better understand how, exactly, the adaptation in question can be employed effectively, become routinized and saved as part of the learned repertoire of responses to everyday, normal, natural troubles[3]. By providing a safe space for gaining experience with such explorations a model supports *learning* as a resilient activity.

No model can do all of these things successfully, because a model by definition requires selecting some elements of a system to be included in its representation, others to be left out, and still others to be extremely simplified. The basis for making these choices depends on the purpose the model is being constructed to address. Therefore, it is important to specifically articulate an objective for modeling before proceeding further. The principal objective of modeling in this thesis is developing explanatory power in two specific areas:

[3] For example, in the ED, the practice of using hallways as spaces to store patients began initially as a radical, desperate reconfiguration. It was first used in multiple ways (*eg*, holding admitted patients *vs* holding newly arrived patients). As experience (including near misses) was gained with the hallway strategy, it was modified (now used only for stable, admitted patients) and ultimately became part of the normal margin for maneuver rather than a unique innovation. So, successful exploration eventually becomes exploitation.

1. Can the model improve our understanding of an event like “free fall”? How did it come about? What might have happened if it could not have been managed?
2. How did a specific adaptation used in “free fall” – a strategy of temporary stopping – actually work? Did it even work, or was the resolution of the crisis only a matter of good fortune? Did it entail additional risks?

I use these two sets of questions to focus model development, guide decisions about what is important to model and at what level of detail, and to inform experiments and interpret their results.

1.3 Notes on models

The modeling approach taken here follows von Bertalanffy’s notion of a general system theory (von Bertalanffy, 1973). In this point of view, a system is viewed not as a collection of components, but rather as a particular type of relation mapping input onto output (Heylighen, Cilliers, & Gershenson, 2007). The internal structure, or the nature of the agents that compose the system, are essentially irrelevant to this way of understanding how it performs that function. The great advantage of this approach is that it makes it possible to establish isomorphisms between systems of different types, *ie*, a way of investigating systems independently of their specific subject domain, by focusing on the pattern of relations among parts rather than the parts themselves. Others have noted when the goal of modeling is to understand qualitative behaviour, then modeling a generic class of systems rather than any single specific system has the additional advantage of making parameter estimation less important in developing useful insights (Forrester, 1985).

The approach is inductive, and informed by the principles of Hollnagel’s *Minimal Modeling Manifesto* (Hollnagel, 1993; Hollnagel, Cacciabue, & Hoc, 1995). Thus, the thesis begins with a series of case studies of practical problems in specific domains. From them it focuses on regularities in the environment and representative ways of functioning across those domains of application. The model development then tries to make as few assumptions as possible and to focus on a core set of essential phenomena, thus resulting in a representation of the kinds of resilient and

non-resilient performance seen in complex sociotechnical systems. This is in keeping with Cilliers' thought that, "It is sometime better to work with a simple model where the limitations are explicit than to work with a complex model that may turn out to be a false friend" (Cilliers, 2001).

These two, related philosophies of modeling lead to the notion that validation of the model is not concerned with establishing the validity of the various internal mechanisms or details, but rather a question of whether "… the variety of the model matches the observed variety of [system] performance" (Hollnagel, 1993). In this regard, it is important to note that many models can be useful even though they are incompletely, weakly, or not at all validated[4]. In particular, they can be useful in summarizing complex data, clarifying ideas, aiding thinking or hypothesizing, or aiding the development of intuition (Hodges, 1991).

People typically bring two common senses to the notion of models (Buck, 1992; Ostrom, Eggertsson, & Calvert, 1990), so I distinguish them here for clarity, as only one is relevant to this thesis. The first sense, the one I do *not* use here, is that of a primarily quantitative and specifically predictive tool that is applicable to a specific situation, whose variables are drawn directly from the domain of application; it is intended to be directly applied to change the world.

The second sense, the one used in this thesis, is more general, can be thought of as perhaps relating a family of more detailed models to one another, and provides a basis for generating more specific hypotheses and / or theories. It identifies the complex system of variables, rule, constraints, *etc*, that affect the performance and control of systems; it is better used to organise and guide understanding (Holling, 1973).

1.4 A caution against reification

Throughout this work, I will often use what are in effect semantic shortcuts – brief phrases that are descriptive of and encapsulate observed behaviours – that might seem to imply something more about the structure or organization of the underlying system of interest. But, I explicitly shrink from the reifying implications of these figures of

[4] For example, most military combat models are unvalidated and unvalidatable, but still may be useful.

speech. For example, in speaking of something like 'short-term memory', I intend nothing more than to recognize the empirical observation that some memories are (often purposely) transient (*eg*, your most recent hotel room number) while others are not (*eg*, your mother's maiden name). The use of such a phrase is a convenient way to describe observations about memory, but should not be further extended to imply that there are separate physiological places / mechanisms / procedures for short term *cf* long term memory. This view is analogous to a view of the "molecular structure" of organic compounds as notational representations of how they react with other compounds. There is no necessity that these "structures" have any physical reality; they are useful in themselves as representations, and if they happen to correspond to the physical structure of a molecule, so much the better. If they do not so correspond, their utility as a representation of reactivity is not decreased.

More specifically, it will often be necessary to refer to some means of control, a way in which a system is able to shift focus, attention, or effort to manage both its work and its adaptations to contingencies. However, although it is often convenient to speak of actors in a system exerting control, this shorthand is not intended to imply that there is a central controller (or set of controllers) directing the system, but should be interpreted as applicable to many varieties of control. Obviously, centralized command and control could be included in this formulation, but it can also be used to refer to distributed control architectures (for example, individual workers in an ED making local decisions and taking local actions that collectively constitute global control); emergent control entirely embedded in the architecture of the system (for example, a slime mold (Dekker, 2011)); or some combination of these extremes (for example, centrally setting goals while allowing local selection of means to attain those goals). Thus when referring to some locus of control, I assume that every system has some means of influencing its actions (else they would be random) but I make no assumption about the specific nature of those means. In addition, this does not imply a strict mechanistic sense of control (*eg*, a linear cause-and-effect chain). In a complex system, actions may control virtually nothing, but influence almost everything (Dekker, 2011); control resides not in the parts, but in the relationships among those parts.

1.5 Overview

This section provides a broad, high-level overview of what will follow. Chapter 2 provides a brief synopsis of the 'free fall' events and uses two existing resilience engineering frameworks to provide a more coherent explanation and representation of the system of the ED and how it responded to and was affected by these events, drawing on accounts that have been published previously (Anders, Woods, Wears, Perry, & Patterson, 2006; Wears, Perry, & McFauls, 2006; Wears, Perry, & Nasca, 2007). (These publications are included in the Appendices). It then presents cases studies from different domains of the behaviour of other systems in crises, and uses them to argue for an underlying similarity among these cases that might be captured in a dynamic model.

Chapter 3 summarizes the resilient responses presented in the case studies by abstracting them into two broad classes, characterized by the labels *exploitation* and *exploration*[5] (Dekker, 2011; Lengnick-Hall & Beck, 2009; James G March, 1991; Maruyama, 1963), and develops the nature of these classes more fully.

One specific adaptation – temporarily stopping the system – will be explored in detail as a special case of the exploration response in Section 3.3. This discusses the special characteristics of stopping in more detail, specifically focused on what it might gain in a system sense, and what characteristics of stopping (when, how long, *etc*) affect its success or failure to help ground the experiments that will be run on the system dynamics model.

Chapter 4 then provides a brief argument supporting the need to develop more dynamic explications of resilience, *ie*, of how resilient performance plays out over time, and proposes to use the system dynamics framework (Sterman, 2000) as a means to approach this goal.

---

[5] These two broad types of adaptations have been variously described under other names. For example, Maruyama called them deviance-reducing and deviance-enhancing actions; Lengnick-Hall refers to them as divergent *vs* convergent forces; others have used change-enabling *vs* change-minimizing, *etc*. Many of the labels that have been applied carry connotations that I wish to avoid here – for example *deviance* is often viewed as a negative, where here I wish to convey a neutral sense that these are merely adaptations, realizing that all adaptations carry the potential for both positive and negative outcomes. March's terms (exploitation *vs* exploration) seem best suited, although I admit that *exploitation* may still carry some negative baggage.

Chapter 5 describes in detail a system dynamics model representing these activities at an abstract, general systems level. It builds a model in a stepwise manner, adding at each step (I hope) just enough complexity to make the model realistic and interesting in its behaviours, but avoiding so fine a level of detail that would make it hard to understand and less generalizable. In this process, rather than first describing the model in detail and then providing an overall justification for its structure and confirming the reasonableness of its performance, I provide these demonstrations at each step, in the hope of making it more comprehensible, reasonable, and convincing to the reader.

Once developed, in Chapter 6 I use this model to investigate the temporal dynamics of resilience, *ie*, the characteristics of the system's ability to respond, specifically with respect to the objectives stated on page 7. For example:

- How long can a response to a shock be sustained?
- What signals are there that a system is exhausting its resilient capacity?
- How does the system "return to normal"? Can it recover at all, and if so, under what conditions and how long does recovery take? If not, how well can it function in some degraded state?
- How does "stopping" work as an emergency adaptation? What sorts of stopping should be tried? When should stopping be invoked? When avoided? How long should it last.

Finally, Chapter 7 will conclude with a discussion of the limitations of this work and an examination of the implications of the findings and areas for future investigation.

French summary of Chapter 1

Ce chapitre commence par un récit d'un événement impliquant une combinaison de surcharge et de capacité réduite dans un service d'accueil et de traitement d'urgences (SAU) de centre hospitalier, qui se trouve au bord de l'effondrement. Ce fut un événement extraordinaire; on prévoit que les SAU fonctionnent à des niveaux élevés 24 x 7 x 365, sans haltes, sans pauses, ni temps d'arrêt, et réussissent en général à cette attente. Ce qui est difficile à exprimer en prose est le sens profond de défaite et d'échec que cet événement produit dans le personnel impliqué. En réponse à cette détresse, je me suis mis à écrire à ce sujet, d'autant plus qu'en discutant de l'épisode dans des entretiens formels et informels avec le personnel impliqué, j'ai découvert un autre événement similaire dans notre SAU qui avait eu lieu environ un mois plus tôt (le 14 novembre 2005). Ceci suggère qu'il se passait quelque chose de plus fondamental, plus qu'une intersection idiosyncrasique de malchances qui n'était pas fortuites. À mesure que je discutais de ces événements avec mes collègues et mes conseillers au cours du mois suivant — décembre 2005 (en particulier avec Richard Cook et David Woods), j'ai commencé à utiliser l'ingénierie de la résilience comme cadre d'interprétation, et j'ai constaté qu'elle m'a permis de donner du sens à ce qui jusque-là n'en avait fait aucun. Ainsi, l'épisode de « chute libre » constitue un exemple motivant pour ce travail - un stimulus à un effort intellectuel plus général. En substance, ces données motivantes forment un «fragment de théorie» et tendent les bras pour l'élaboration et le développement d'une vraie théorie (Davis, Eisenhardt, & Bingham, 2007).

La résilience peut être définie comme «la capacité intrinsèque d'un système d'ajuster son fonctionnement avant, pendant ou après les changements ou les perturbations, afin de pouvoir soutenir les opérations nécessaires dans des conditions à la fois attendues et inattendues» (Hollnagel, 2011). Cette thèse étend les modèles théoriques actuels de la résilience (Hollnagel & Sundström, 2006; Woods & Wreathall, 2008; Woods, Wreathall, & Anders, 2006) en développant des modèles du système de simulation dynamique (Sterman, 2000) pour représenter des concepts théoriques et des relations sur la performance résiliente et sa dynamique. L'idée fondamentale est que la stabilité résiliente que nous voyons dans des systèmes complexes est le résultat

d'un équilibre dynamique. Cette thèse utilise la dynamique de modélisation et de simulation du système pour explorer les propriétés émergentes d'une théorie de la dynamique de la résilience.

La modélisation peut fournir plusieurs avantages généraux: des explications des phénomènes observés, l'identification des voies possibles qu'un système pourrait traverser; l'identification des indicateurs potentiels de perte de marge de manœuvre (Pariès, 2011; Stephens, 2010; Stephens, Woods, Branlat, & s'habille, 2011), ou le rôle de la synchronisation des adaptations, la capacité de tester sans risque les réponses possibles, l'occasion pour la pratique de contrôler les situations dangereuses, ou une aide dans l'apprentissage de l'organisation. Cette thèse se concentre sur le premier avantage, le pouvoir explicatif, dans deux domaines spécifiques:

1. Le modèle peut-il améliorer notre compréhension de la façon dont une «chute libre» a lieu?
2. Comment les adaptations d'arrêter le système ont-elles réussi — à supposer que celles-ci ont réussi?

L'approche de modélisation qu'on a pris était basée sur la notion de von Bertalanffy d'une théorie général du système (von Bertalanffy, 1973), éclairé par les principes du Manifeste Hollnagel de modélisation minimale (Hollnagel, 1993; Hollnagel, Cacciabue, & Hoc, 1995). Ainsi, cette approche comprend le moins d'hypothèses possibles, et n'est pas un outil quantificatif ni spécifiquement prédictif, mais plutôt un moyen d'identifier le système complexe de variables, de règles, de contraintes, etc., qui affectent les performances et le contrôle des systèmes, afin de comprendre l'organisation et l'orientation (Holling, 1973). Le chapitre se termine en établissant un guide de ce qui va suivre.

Chapitre 2 fournira un résumé plus détaillé de la «chute libre» des événements, et présenter des études de cas à partir d'autres documents pour soutenir l'idée d'une similitude fondamentale parmi ces cas, il concerne également leur modèles actuels de la résilience.

Le chapitre 3 résume les réponses résilientes dans les études de cas, les abstrayant en deux grandes classes, appelées l'exploitation et l'exploration. Le chapitre 3 traite explicitement une adaptation spécifique de l'exploration - l'arrêt du système - en termes de ce qui pourrait être acquis (ou risqué) en l'arrêtant.

Chapitre 4 donne une brève argumentation de la nécessité d'une compréhension dynamique de la résilience.

Le chapitre 5 expose le modèle développé pour ce travail de manière progressive, en commençant par un modèle extrêmement simple, puis en ajoutant juste assez de complexité, pour rendre le modèle plus intéressant et plus réaliste dans ses comportements.

Le chapitre 6 utilise ensuite ce modèle dans une série d'expériences de simulation pour explorer la nature de la performance résiliente et les facteurs qui la favorisent ou la dégradent.

Et enfin, le chapitre 7 conclut par une discussion des limites de ce travail, un examen de ses implications, et des domaines d'investigation future.

Chapter 2. Models of resilience

The "free fall" events, and a related examination of ED work under less demanding, "normal" circumstances, are given in greater detail in a series of papers and book chapters included in Appendix 3 (Wears, Perry, Anders, & Woods, 2008; Wears, *et al.*, 2006; Wears, Perry, & McFauls, 2007; Wears, Perry, & Nasca, 2007). Here I give only a brief background and synopsis of these observations to aid in understanding the subsequent analysis.

2.1 Background

Essentially, both events involved crises of over-crowding in the ED. ED / hospital crowding is a serious international problem that has grown worse over the past 30 years. The problem first became apparent in US EDs in the 1980s, and was thought to be of crisis proportions by the end of that decade. The American College of Emergency Physicians issued a position statement (1990a) and several policy recommendations (1990b) on what was then called "emergency department overcrowding" in 1990, but the problem only continued to grow (R. W. Derlet & Richards, 2000; Goldberg, 2000; A. L. Kellermann, 2000; Zwemer, 2000). Eleven years later, in 2001, the Society for Academic Emergency Medicine made crowding the theme of its yearly Consensus Conference; entitled *The Unraveling Safety Net*, the Conference resulted in the dedication of an entire issue of the Society's journal, *Academic Emergency Medicine*, to a group of papers on the crowding problem (Adams & Biros, 2001; Baer, Pasternack, & Zwemer, 2001; R. Derlet, Richards, & Kravitz, 2001; Gordon, Billings, Asplin, & Rhodes, 2001; Kelen, Scheulen, & Hill, 2001; Reeder & Garrison, 2001; Schneider *et al.*, 2001; Schull, Szalai, Schwartz, & Redelmeier, 2001). Despite this attention, crowding has only gotten worse in the ensuing years (Arthur L. Kellermann, 2006; US General Accounting Office, 2003), culminating in a 2006 Institute of Medicine report that warned that the emergency care system in the US was on the verge of total breakdown (Institute of Medicine, 2006).

The crowding problem at Shands Jacksonville paralleled these national patterns. As the problem grew more severe, the ED adapted by routinely using non-standard

spaces (aisles, chairs, hallways) to temporarily manage patients during crowded periods. Chairs near the nursing station (rather than stretchers in treatment rooms) came into routine use in the 1990s. The practice of using additional stretchers in the aisles of treatment areas started in the 2000s, as did the practice of holding stable, admitted patients in hallways just outside the treatment areas when no beds were available on the routine hospital wards. Because ED crowding had often created problems with ambulances being diverted to other hospitals, in 2002 the city's Department of Public Safety banned the practice of ambulance diversion for crowding. Finally, in early 2004, the ED decided to reserve one of its five major treatment areas (constituting 28 beds) solely for holding admitted patients (called "boarders") and withdraw from this space, essentially shrinking into a 20% smaller footprint in which to treat the same volume of ED patients. All these adaptations had been long-standing and were viewed as generally successful at the time of the "free fall" events. Although they were not explicitly framed as such by workers in the domain, these adaptations would be easily framed as efficiency-thoroughness tradeoffs (Hollnagel, 2009). By giving up some of the resources (space, staff time, *etc*) nominally assigned to new incoming patients, the ED is able to manage a much larger load of patients (both new incoming patients and the boarders) than it was designed or staffed) to handle. It should also be noted that, particularly in their beginning, many of these adaptations were officially viewed as deviant, although often tacitly accepted (for example, treating patients in chairs and temporary spaces rather than having them lay on stretchers, or leaving stretcher patients in hallways).

2.2 Synopsis

In the interest of space, I will describe only one of the "free fall" events here; both played out in quite similar ways, and more details are provided in the papers in Appendix 3.

The ED was severely, but not unusually, crowded at the onset of the episode. The 21-bed critical care unit was full; its four bed resuscitation area was fully occupied with patients on ventilators, four patients were being managed on stretchers in the aisles, and another seven in the hallway just outside the unit. During the first few hours of the evening shift (roughly 1500 to 1700), the combination of the arrival of several

critically ill patients in rapid succession and the "normal" inflow of self-referred patients with conditions requiring the critical care area led in a few hours to a situation of total gridlock. Noise levels rose, making it difficult to communicate. Physical congestion was a major problem, making it difficult to move about in the treatment area; this was a doubly significant problem because certain patients require certain equipment, so that either patients or equipment need to be moved rapidly and on short notice. In addition, the level of congestion precluded use of one of the ED's more characteristic adaptive mechanisms, bringing in more staff. In this situation, adding more people to a confined, already physically congested workplace would have only made the situation worse. As the situation deteriorated, the staff eventually "lost the bubble" – lost their mental picture of the numbers, types, or problems of the patients for whom they were responsible (Roberts, 1989). This was an unexampled event in the experience of the staff, and a sense of control was only regained after stopping all attempts to provide anything other than immediately life-saving care, and then systematically enumerating the patients who were present, prioritizing their problems and reassigning staff to resume their care; essentially, by stopping the system and then gradually restarting it.

Although all concerned agree these two episodes were periods of high risk, as far as is known, no patient suffered harm related to these events. (One patient did suffer a serious adverse event, but the causal links between the "free fall" episode and her outcome are both unclear and contestable).

The crowding problem and the "free fall" episodes that resulted represents classic "going sour" incidents, with a characteristic two-phase signature (Woods, Dekker, Cook, Johannesen, & Sarter, 2010). In the first phase, there is a slow degradation of the monitored process, a gradual falling off over a period of time. Because of the slowness of this process, it represents a very soft signal (if it is interpreted as a signal at all) that the potential for loss of control is growing. That is, it can be difficult to distinguish a major challenge, partially compensated for, from a minor disturbance that can be safely ignored or expected to resolve on its own. Eventually, if operators do not intervene in a timely and effective way, a rapid collapse occurs. This reinforces Woods' notion that the critical difference between a major and minor

disturbance is not their symptoms, but rather the force with which they must be countered (Woods, *et al.*, 2010).

2.3 State space model

The resilience state space model (Hollnagel & Sundström, 2006) provides a compact way to summarize the state of the ED as it progressed into and out of these crises (see Figure 1). This model describes a system as being in one of several states, where a state is defined as "any well-defined condition or property that can be recognized if it occurs again" (Ashby, 1957). Typically, these states would include a state of normal functioning (where the system operates as intended in a reliable way), a state of regular reduced functioning (such as nights, weekends, holidays, or scheduled downtimes), and a state of irregular reduced functioning (such as equipment failures, unexpected staff absences, *etc*). Unfortunately, the available states must also necessarily include states of disturbed functioning (corresponding to severe, unexampled, or catastrophic situations), and a state of repair (where normal production and operations cease). The model also describes transitions among states, which for all but the scheduled transitions between normal and regular reduced functioning, are typically associated with losing or regaining control.

Based on work by Voß et *al*, I have modified the original diagram for the resilience state-space model to more clearly separate states that are commonly experienced from those that are rarely or never experienced (Voß, Procter, Slack, Hartswood, & Rouncefield, 2006). The former represents situations of expectable, "normal, natural troubles[6]" whose solution is readily available to members and their normal working practices; *ie*, they require nothing more than the "usual solutions" that are an ordinary part of day-to-day work. The dotted line in Figure 1 labeled *Horizon of Tractability* separates these states from states representing more severe disruptions and requiring solutions that are relatively more extreme, novel, creative or untried. As suggested by the dotted, wavy line, this boundary is both uncertain and shifting, so that its precise location at any given time is unknowable. A system crosses the ‘horizon of

[6] I use the phrase "normal, natural troubles" in its conventional sociological sense of meaning "in accord with the taken-for-granted way of doing things" (Garfinkel, 1967).

tractability' when it moves from a manageable, familiar if disrupted state to a novel or difficult-to-manage one due to unexpected developments, or because its functioning and status are not longer well understood.

In the language of this model, these episodes began in a state of 'regular reduced functioning' – this was not 'normal functioning', at least in the normative sense, because the ED was chronically decompensated due to crowding as noted above. It remained operable, but at a reduced level of functioning and a reduced margin of safety. This state is above the 'horizon of tractability' and so indicates workers are responding to the "normal, natural troubles" of work by relying on familiar and commonly used adaptations (Anders, *et al.*, 2006). As events unfolded, increasing demand combined with limited or decreasing resources to produce a loss of control and a transition into the state of disturbed functioning. While in this state, progressively more extreme and novel adaptations were required to maintain any semblance of production; the specific adaptation observed in these events, and in more nearly normal operations, are discussed in detail in Section 3.2.

As the evening wore on, the state of the system remained in 'disturbed functioning' for a long period of time and in fact progressively worsened, until finally the novel decision was made to stop all ordinary care activities, exhaustively enumerate and identify patients and their problems, and then to reprioritize work, reassign staff, and thus resume ED operations. In the state-space representation, this corresponds to a retreat to the repair state, followed by a resumption of operations in either the regular or irregular reduced functioning state.

It is interesting to note that one of the lessons learned from this experience was that the ED stayed for too long in the disturbed function state, hoping against hope for a recovery transition, when moving more quickly to the repair state was indicated. The delay was likely due to three factors. First, the severity and duration of the disturbed functioning state was unprecedented in the collective experience of the ED staff, so there was no prior experience with this solution. Second, these episodes developed insidiously, without a clear and unambiguous external trigger (*eg*, a plane crash, or tornado, *etc*) that could have signaled the need for either unprecedented action, or for triggering little-used but "on the books" responses such as the hospital's disaster plan.

Figure 1. Resilience state space diagram (modified from Hollnagel & Sundström, 2006).

My feeling (and this has been corroborated in interviews with staff involved) is that if faced with such a situation again, we would do better to invoke the retreat into repair transition sooner rather than later as this would allow earlier restoration of a more stable operating state and expose fewer patients to risk over a smaller period of time. Essentially, a small trade-off of thoroughness for efficiency, if done early enough, could have avoided the much larger and more risky trade-off that ultimately occurred. This is, in effect, a trade-off among tradeoffs (small, frequent early trade-offs vs large, less common ones).

This complexity of tradeoffs does not seem to have been much explored. Hoffman and Woods have argued that five fundamental tradeoffs along the dimensions of ecology, cognizance, perspectivity, responsibility, and effectivity characterize complex adaptive systems performing macrocognitive work (Hoffman & Woods, 2011); any strategy for operations will involve tradeoffs on one or more of these dimensions. Thus choosing any specific strategy itself involves a multilevel tradeoff by forcing a choice among sets of different tradeoffs.

This highlights two potentially valuable areas of investigation that will guide investigations of the model. First, I will explore the potential for signals that a state transition is imminent or would be beneficial; such signals might allow workers to take steps to either bring it about or forestall it, depending on the valence of the transitioned-to state (desirable or undesirable). Second, the model might be able to suggest ways to evaluate the trade-off among trade-offs dilemma, by helping operators estimate the likely effect of such adaptations.

2.4 Stress-strain model

Woods and Wreathall have suggested an alternate representation of resilience, based on an analogy to the stress-strain curves common in materials science, which is illustrated in Figure 2 (Woods & Wreathall, 2008; Woods, *et al.*, 2006). In this representation, the y-axis represents demand on the system and the x-axis represents how the organization changes ("stretches") in response to that demand. (Note that this reverses the typical arrangement of independent and dependent variables in analysis, but follows the convention of stress-strain plots in materials science).

Stress-strain plots typically are divided into two regions: an elastic region where the material stretches uniformly under increasing load and a plastic region where the material begins to stretch uniformly until distortions and gaps accumulate and a fracture or failure point is reached. In the elastic region, the response to increasing demands is proportional while in the plastic region the material cannot stretch completely to meet the demand. In Figure 2, the straight line marked ‘uniform’ on the left of the graphic represents the elastic region of the system. Here, the normal, natural responses to normal, natural troubles work to allow the system to respond smoothly to demand. When this routine adaptive capacity is exceeded, then the system enters the plastic, or ‘extra’ region. Here, routine, first-order responses are no longer adequate to meet demand and to avoid gaps and failures, additional extra adaptations requiring new work, new resources or new strategies come into play to allow demand to be at least partially met, albeit at greater cost (in resources, effort, speed, and/or quality). Progressively increasing demand leads to new adaptations and reconfigurations (deformations) until ultimately adaptive capacity is exhausted and the system fails (the material fractures) – or, the system re-organises and functions in

a new mode. This restructuring creates new (perhaps smaller) areas of response, with new uniform and plastic regions.

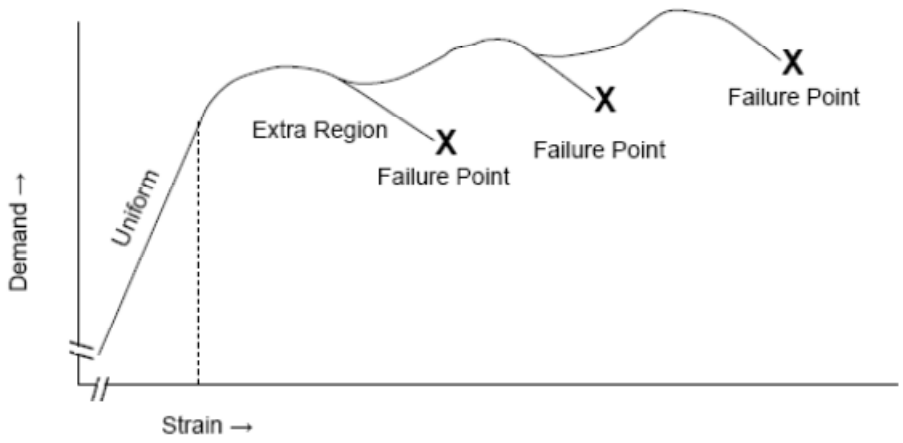


Figure 2. Stress-strain representation of resilience (from Woods *et al*, 2006).

The stress-strain analogy is appealing because it suggests possible empirical application. If reliable and valid measures of demand and resource investment were available, then several measures might be useful. The slope of the elastic region would represent the normal performance capacity of the system. The level of demand at which elasticity is lost would be its maximal normally tolerated demand. Since many of the adjustments that improve performance in the uniform region paradoxically increase fragility or brittleness in the extra region (*ie*, in situations of excessive, novel, or unanticipated demand) (Carlson & Doyle, 2000, 2002; Csete & Doyle, 2002; Zhou, Carlson, & Doyle, 2005), the ability to identify the point of transition between the elastic and plastic regions could be useful in avoiding problems of over-optimisation, where organizations adapt so perfectly to some set of circumstances that they cannot function when those circumstances change. The average slope in the multiple deformation region reflects the adaptive capacity of the system. And of course, the level of demand at the point of failure might be estimable.

2.5 Parallels to other domains

The ‘free fall’ events serve as case examples, or existence proofs, that there is a real problem with the stability of the ED as a system. Qualitatively similar trajectories have been noted in other systems; some brief examples of these follow. The purpose

of these examples is to demonstrate that the problems and behaviours identified in the ED case studies are not unique to the ED, or to healthcare, but are representative of a more general class of issues in many complex systems. Although the specifics differ by domain, there is an underlying similarity in these events which lends support to the notion of developing a generalized, abstract model that might usefully focus attention on essential rather than idiosyncratic, domain- and context-specific aspects of these systems.

Each of the three examples that follow might be easily represented by either of the state-space or the stress-strain analogues. However, like the 'free fall' events, they all also show dynamic properties that are difficult to capture in these representations, thus supporting the need for a more dynamic scheme.

2.5.1 California electric power transmission grid

Paul Schulman and his research group studied how the organizations responsible for the reliable distribution of electrical power responded to the restructuring of the energy market in California in 1996 (Roe & Schulman, 2008; Schulman & Roe, 2007; Schulman, Roe, Eeten, & Bruijne, 2004). The bulk of their data were gathered from the control room operations of the California Independent Systems Operator (CISO, the entity charged with managing the grid, and separated from the generation, purchase, or sale of power) in 1999 through 2001. During this time, because of market manipulations later found to be illegal and unethical, but also because of other unintended and unanticipated consequences of deregulation, California experienced an unprecedented crisis in the availability and pricing of electrical power (Congressional Budget Office, 2001; Sweeney, 2002). Operation of the power grid is both complexly interactive (in the sense that there are non-linearities between stimuli and responses, and that effects of local changes can be felt remotely) and tightly coupled (in that there is little capacity for buffering; power cannot be stored, and it is critical to match available generation with current demand on the dimensions of power (voltage and current), frequency, and phase. The CISO balances load and generation in real time by a repertoire of responses and options in the face of unpredictable or uncontrollable system instability produced either within the network (*eg*, by generators acting in a strategic fashion) and from outside the network through

its open system features (*eg*, high ambient temperatures). 'Load' is the demand for electricity and 'generation' is the electricity to meet that load; these must be made equal within brief periods of time, otherwise service delivery is interrupted as the grid physically fails or collapses.

Schulman notes that, despite the widespread media coverage of the crisis in the availability and pricing of electricity, what was remarkable about the California energy crisis was not how disruptive it was, but rather how system operators managed to "keep the lights on" in a complexly interactive, tightly coupled system under un-exampled conditions[7]. They used both the Berkeley group's high reliability organization framing (La Porte, 1996; Roberts, 1989; Rochlin, La Porte, & Roberts, 1987; Schulman, 1993a; Weick, 1987; Weick, Sutcliffe, & Obstfeld, 1998) and Perrow's normal accident theory (Perrow, 1984, 1994) framing in their analysis; they found ways in which both were insufficient, and proposed their own extensions.

They summarized their observations by noting that control room operators shifted among 4 modes of operation, based on the relative balance between equifinality (the availability of multiple options to a manage or solve a problem) and network instability (the presence of rapid, uncontrollable changes in demand or generation, or uncontrollable external factors such as weather.) Table 1 provides their representation of these modes, based on high *vs* low instability and high *vs* low option variety. We describe these modes in order from the most stable and most preferred to the least stable and least desired, in order of progressive loss of control (*cf* pg 18); this ordering moves anti-clockwise from the just-in-case to the just-this-way cell.

Just-in-case performance occurs in situations of abundant resource and organizational slack (Schulman, 1993b). The system is close to static equilibrium, there is little unpredictability or uncontrollability, resources are abundant, and multiple options for control inputs, should they be needed, are readily available. This represents the state of "normal operations", so seldom achieved in the real world, but on which most

[7] For example, during the peak of the crisis in 2001, the net effect of the rolling blackouts that were used in the most critical circumstances was equivalent to less than 1 hour's outage for each of the 11.5 million households in the state.

official operational policies and procedures are based. *Ie*, it is the operational state most common in theory.

Table 1. Performance modes for the CISO control room (from Schulman *et al*, 2004)

		System instability	
		High	Low
Network option variety	High	Just-in-time	Just-in-case
	Low	Just-for-now	Just-this-way

Just-in-time performance occurs in situations where instability is high, but the options available to operators are similarly high. It is a state of dynamic equilibrium, where the system is maintained in a stable state by continuous control inputs from the operators, who are not severely restricted in the suite of available options from which they can choose. It represents the state that is most likely common among a large number of complex systems that generally achieve high levels of performance via the continuous supervision and manipulation by those in charge of it. *Ie*, it is the operational state most common in practice.

Just-for-now performance occurs in situations where control options are few, but unfortunately, instability is high. Here, operators begin to move into crisis management, and have no illusion that they are in control, but feel that they are being driven by events, exemplified by their use of the term 'firefighting'. A major concern here is the risk of deviance-amplification (Maruyama, 1963); that small changes in inputs or contextual factors can ramify rapidly throughout the system. From the standpoint of resilience, this situation is untenable; the buffering resources and margin for maneuver are close to exhaustion, and the risk of total collapse is palpable.

Just-this-way performance is the last resort in regaining control. System stability is regained by the execution of a limited option, that of some form of shut-down. In Schulman's situation, this was achieved by the assertion of prescriptive, command-

and-control authority over generation and distribution in the form of blackouts. In our more general formulation, this would be a form of the stopping response.

For our purposes, we can recast these 4 modes into our two large classes of exploitation (margin maintaining) and exploration (reconfiguration). In just-in-case performance, there is plenty of margin for maneuver and no need to reconfigure. Just-in-time performance uses additional resources (*eg*, buffers) to maintain the margin and normal operations. Just-this-once performance begins to slip into novel reconfigurations – modes of working that would not normally be considered, but are tolerated "just this once" in order to achieve higher level goals. And just-this-way performance is the ultimate in reconfiguration, the strategy of stopping the system (in whole or in part) in order to regain control.

2.5.2 IT system crash – situational on fundamental surprise

Shortly before midnight on a Monday evening, a large urban academic medical center suffered a major IT system crash which disabled virtually all IT functionality for the entire campus and regional outpatient clinics (Wears, 2010; Wears & Webb, 2011). The outage persisted for 67 hours, and forced the cancelation of all elective procedures on Wednesday and Thursday, and diversion of all ambulance traffic to other hospitals. (52 major procedures and numerous minor procedures such as colonoscopies were cancelled; at least 70 incoming ambulance cases were diverted to other hospitals). There were substantial (4 to 6 hour) delays in both order and obtaining laboratory and radiology results which severely impacted ongoing clinical work. A previous risk analysis had estimated direct costs for complete downtime at $56,000 per hour, so the total direct cost (not including lost revenue from cancelled cases or diverted patients) is likely close to $4 million. As far as is known, no patients were injured during this event, and no previously stored data were lost.

The triggering event was found to be a hardware failure in a network component. This interacted with the unrecognized presence of modules from an incompletely aborted (and ironically named) "high availability computing" project some years previous such that the system could not be restarted. The restart failure could not be corrected initially because of a second, independent hardware failure in an exception

processer. Once this was identified and replaced, the system still could not be restarted because unbeknownst to the IT staff, the permissions controlling modifications to the start-up files and scripts had been modified during the high availability project, so that no one in IT was able to edit them to make corrections and thus re-start the system.

After the brief initial delay, the hospital was able to quickly reorganize in multiple ways to keep essential services operating in at least some fashion for the duration. Adaptations included exploitation of existing resources or buffers; and exploration of novel, untried ways of working.

For example, adaptations of exploitation included deferring elective cases or encouraging early discharges of patients who were improving. These adaptations were limited in scope, because the extent of the problem was not realized until Tuesday's list of elective cases was well underway. Similarly, plans for early discharges were stymied by the slow delivery of laboratory and imaging results due to the system outage; physicians were reluctant to discharge patients when such results were still pending.

Several adaptations of exploration were also invoked. An incident command centre was set up. Because the geographic area experiences frequent hurricanes, the incident command system was well-rehearsed and familiar, so it was adapted to manage a different type of threat. A similar novel use of a well-rehearsed technique was used to mitigate the loss of medical record numbers while the system was down. The ED had been planning to implement a 'quick registration' method, where only basic patient information is obtained initially to permit earlier orders and treatment, and the registration process is completed at a later time. The IT failure prevented complete registration but was thought to have left the capability for quick registration. Because this method was very close to implementation, so it was pressed into service. However, its application in this setting uncovered a problem, in that different organisational units used the same variable to represent different information; this resulted in several patients in getting "lost" in the system. This failure led to an alternative, the use of the mass casualty incident (MCI) system.

In anticipation of occasional mass casualty incidents in which the numbers of arriving patients would too rapidly exceed the ability of the registration system to record their basic information and assign them identifying medical record numbers, the organization maintained a separate set of so-called MCI numbers and armbands. The MCI system is normally used for situations of high demand that exceeds available resources (*eg*, an explosion, or a plane crash). However, it is formally designed to accommodate any mismatch between demand and available resources. In this case, demand was normal to low, but resources were much lower, so the MCI system was used to identify and track patients and their orders, medication, procedures, and results, with a plan to marry the MCI information to formal medical record numbers after the incident had been resolved.

The most novel adaptation of exploration included rescheduling financial staff (who now had nothing to do, since no bills could be produced or charges recorded) to use them as runners to move orders, materials, and results around the organization that had previously been transmitted electronically.

2.5.3 Rule-violating use of a tightly controlled drug

An extremely agitated and violent young man was brought into the emergency department (ED) after having attacked a police officer following a minor traffic accident due to erratic driving (Gilardi, Guglielmetti, Perry, Pravettoni, & Wears, 2009). The police had used a Taser® (an electric shock device) multiple times to try to control him. No other history was available; thus, there was a great deal of uncertainty about the etiology of his condition, and the space of possibilities included at least drug overdose, head injury, or severe psychosis.

He had been shackled by the police in an unusual position; his wrists were chained together behind his back; his ankles were chained together; and his wrists were chained to his ankles, rendering him face down, back arched on the stretcher. This positioning made the ED staff very uncomfortable, because if he should deteriorate, they would be unable to turn him from the prone to supine position in order to protect

or establish an airway[8]. This discomfort was exacerbated by the fact Tasers® have been suspected of causing sudden, unexpected cardiac arrest, and that only weeks prior, a patient who had been Tasered by the police had suffered an unexpected cardiac arrest in this ED.

This situation represented a combination of *thought-of* and *un-thought-of* conditions (Cuvelier & Falzon, 2011). Violent, agitated patients are not unusual in EDs, and there is a well-defined protocol for managing them, the "rapid sedation protocol." However, managing an airway with a patient restrained in this position is an unexampled event.

The ED staff began to follow their rapid sedation protocol, but rapidly became dissatisfied with it as a solution. The protocol involves giving small doses of sedating medications intravenously at frequent (3 to 5 minute) intervals until the desired effect is achieved; this is to avoid the risk of over-sedation and its adverse consequences. However, it often takes 20 to 40 minutes to achieve sedation, and this patient remained so violently agitated after the first doses that the chances of losing the intravenous line, or experiencing some new complication prior to achieving control seemed high. In addition, the option of temporarily removing some of his restraints to allow repositioning in the supine position seemed particularly risky for both patient and staff.

After a brief discussion among the team (all with long ED experience), a novel alternative was adopted. They abandoned the established protocol and decided instead to use another drug, propofol, to reach a state in which any future untoward events might be more controllable. Propofol is a powerful sedative agent whose action is both rapid (within 5 – 40 seconds) and brief (3 – 5 minutes). However, its use is strictly controlled in the organization because it is also dangerous; it can rapidly stop breathing and severely lower blood pressure[9]. Using propofol here violated

---

[8] Ensuring an airway is the first and highest priority in a medical emergency; inability to ensure an airway leads to death or severe brain injury in a matter of minutes. It is difficult to express the degree of discomfort this patient's position provokes in an emergency care provider.

[9] It is also strictly controlled for political reasons; propofol was for a long time restricted to anesthetists, some of whom viewed extending its use to ED physicians as encroaching on their "turf".

organizational protocols for propofol in three ways: in the indication (no painful procedure was planned); in the preparation (no vital signs could be obtained, and no pre-anesthetic assessment could be done); and in the knowledge base (there is neither much experience nor published material on the interaction of propofol with street drugs). But, this strategy was (fortunately) successful. The patient was moved to a resuscitation bed and the team set up for an emergency airway; the patient was then given intravenous propofol and when he lost consciousness, he was unshackled, repositioned and re-restrained, and then allowed to wake up.

Again, this vignette contains strategies of exploration and exploitation. The work team anticipated the failure of the standard plan (Branlat, Anders, Woods, & Patterson, 2008), and switched to a novel plan. The new plan involved the application of procedures well-rehearsed in other contexts (*ie*, using propofol for brief sedation for painful procedures, such as fracture reduction; and setting up for emergency airway management) to a novel setting; this is essentially an enactment of the concept of the adjacent possible (Johnson, 2010; Kauffman, 1995). In addition, it involved accepting a brief period of high risk to gain better ability to control potential future events – action in the present to help maintain control in a possible future, *ie*, a tradeoff of current risk for future safety (Hoffman & Woods, 2011).

French summary of Chapter 2

Ce chapitre fournit un résumé plus détaillé de l'une des deux "chutes libres" d'événements. (Une seule est présentée en détail, les deux épisodes sont décrits avec plus de détails dans les documents inclus dans l'appendice 3.

Ces événements furent tous les deux des crises de surpeuplement au SAU. Le surpeuplement dans les SAU aux États-Unis augmentait depuis 25ans, malgré une reconnaissance du problème et de multiples tentatives de le contrôler. Le jour en question, l'SAU était sévèrement, mais pas exceptionnellement bondé. Ses 21 unités de soins intensifs furent complètes. 11 malades supplémentaires se trouvèrent gérés tant bien que mal, soit dans les allées ou dans le couloir juste à l'extérieur du SAU. Au début du poste du soir, une série de gravement malades se présentèrent en succession rapide; en combinaison avec l'afflux normal de malades graves (par exemple, de douleurs thoraciques, de pneumonie, d'asthme). Ceci provoqua une congestion physique grave et, finalement, le personnel perdit son image mentale des nombres et des types de malades, dont on était responsable. Ce fut un événement sans exemple et on ne put revenir à la raison qu'après l'arrêt de toutes tentatives de fournir quoi que ce soit hormis la réanimation. En suite on énuméra systématiquement les malades présents, puis on établit des priorités, puis on réaffecta le personnel aux soins. Essentiellement, on arrêta le système, avant de le redémarrer progressivement. Bien qu'il parût clair que ces deux épisodes furent de risque élevé, autant qu'on le sût, aucun malade ne subit de préjudice directement lié à ces événements.

On raconte ensuite les événements dans ce cas à deux modèles de résilience généralement acceptés et synergiques : le modèle «espace-état » (Hollnagel & Sundström, 2006), et le modèle «contrainte-déformation» (Woods & Wreathall, 2008;. Woods et al,2006). Le modèle espace-état envisage un système d'exploitation dans l'un de plusieurs états (par exemple, le fonctionnement normal, réduit, ou perturbé), et de faire les transitions entre ces états en réponse à des contraintes qui provoquent une perte du contrôle ou des adaptations visant à le rétablir.

De même, le modèle de contrainte-déformation (basé sur une analogie avec la science des matériaux), décrit comment un système se modifie à mesure que les contraintes

augmentent. D'abord il s’étend progressivement et élastiquement en réponse à la demande; c’est-à-dire des adaptations routinières à des problèmes routiniers. Ainsi que l’effort s’accroît, il s'écarte d’une région «élastique» à une «plastique», dans laquelle il «fausse» — des adaptations supplémentaires nécessitant des travaux, de nouvelles ressources, ou des stratégies nécessaires à remplir au moins partiellement la demande. Des demandes qui mènent progressivement à de nouvelles adaptations et à des reconfigurations jusqu'à ce que le système échoue finalement (les fractures du matériel).

On présente ensuite trois études de cas provenant d'autres domaines (transport d'énergie électrique, technologie de l'information et de gestion de combat des malades violents) pour étayer l'idée que, à un certain niveau d'abstraction, les réponses du système soient fortement similaires, et qui tombent dans les modèles d'exploitation des capacités et des ressources bien comprises (par exemple, la marge de manœuvre), ou l'exploration des capacités et des réponses neuves et jusque-là inexplorées.

Chapter 3. Abstraction

Although the examples just related and the 'free fall' cases differ greatly in their specifics, I propose that they have an underlying unity in the types of responses to challenges they demonstrate, and to simplify their complexity not by reduction, but rather by abstraction.

3.1 Classes of adaptations

The specific responses seen in these cases can be usefully represented by two broad classes, characterized by the labels *exploitation* and *exploration* (Cuvelier & Falzon, 2011; Fu, 2007; Lengnick-Hall & Beck, 2009; James G March, 1991; Maruyama, 1963; Pariès, 2011; Piaget, 1967). This classification is not meant to be either perfectly unambiguous or comprehensive, but merely useful for the purpose of developing a representation of resilience dynamics.

Exploiting adaptations are a set of generally familiar responses to generally familiar problems; they serve as a sort of internal momentum, a homeostasis, in that they allow the system to absorb some shocks while continuing to function at some level in its current state. Exploring adaptations, on the other hand, are fundamentally major and generally novel reconfigurations of the system, substantive changes that allow function to continue by moving to some new state[10]. In addition to allowing continued operation in the face of external or internal challenges, they support organizational learning by experimenting with new ways of operation.

Others have advanced the notion of these two broad sorts of response as well. Maruyama has pointed out that control engineering has tended to focus on negative-feedback-based, "deviance reducing" processes, but that in many real world systems, positive-feedback-based, "deviance enhancing" processes also contribute importantly to overall development or performance (Maruyama, 1963). He went further to note that although these processes operate simultaneously, one or the other tends to dominate at any given time, so that the interplay between the two types of process

[10] A classic example is Weick's description of the life-saving "escape fire" strategy used by Wag Dodge when his crew was over-run by a wildfire (Weick, 1993).

was just as important in explaining behaviour as the specifics of the processes themselves, so that they might legitimately be called "mutually causal processes." A difficulty with Maruyama's nomenclature is his use of the term "deviance", which is generally interpreted pejoratively, although a close reading of his work confirms this was not his intent.

Similarly, Lengnick-Hall noted the importance of having two different sorts of responses which she called divergent and convergent forces, and in particular noted that different circumstances require different responses (Lengnick-Hall & Beck, 2009). In her view, in "moderately unsettled" situations (think of "normal, natural troubles"), complexity-reducing, convergent, Taylorist, pre-planned, distant supervision responses were most appropriate, while in "highly turbulent" circumstances (think of unexampled, or off-design-base circumstances), complexity embracing, divergent, local action, novel, even counter-intuitive responses worked best.

Piaget noted a similar dichotomy of responses, which he termed assimilation and accommodation (Piaget, 1967). He thought of assimilation as basically homeostatic, while accommodation required self-modification, typically triggered by cognitive dissonance (Pariès, 2011). A similar dichotomy of response types has been described in resilient decision-making in anesthesia (Cuvelier & Falzon, 2011). Lundberg has also noted a distinction between the ability to cope with known disturbances (stability) and the ability to cope with irregular or unexampled events (resilience) (Lundberg & Johannson, 2006). His formulation mirrors that of Hollings, who is thought to have first use the term 'resilience' in a sense close to how it is used in resilience engineering, in his framing of the dynamic of resilience and stability in ecological systems (Holling, 1973). And March has expanded on this theme and proposed (in my view) the most useful labeling, contrasting exploration and exploitation as complementary strategies for enhancing organizational resilience (James G March, 1991).

What unites all these schemes, beyond the obvious correspondences in their parts, is the notion that although the two types of responses exist in a sort of dynamic tension, both are necessary for long term success (Carlson & Doyle, 2000, 2002; Fu, 2007);

one must be balanced against the other in a fundamental tradeoff (Hoffman & Woods, 2011). A system too heavily dependent on exploitation – too well adapted to its current environment – would be unable to reconfigure itself and change in response to a change in circumstances. Conversely, a system too dependent on exploration would be too inefficient to prosper. This constitutes an optimality – resilience tradeoff (Woods & Branlat, 2011b): to be successful in the long term, complex systems must always be partially "de-tuned", such that a (sufficiently small) in insufficiency in exploitation responses leads to exploration responses, which can lead to new behaviours (Lundberg & Johannson, 2006). As these are successful and become learned, they enhance the repertoire of exploitation; but they also expose the system to still newer challenges by changing the internal and external environment. A particular goal of this work is to examine the possibility of identifying synergistic or antagonistic factors affecting which type of adaptation is dominant, and indicators of impending shifts (desirable or undesirable) between exploitation and exploration.

3.2 Observed adaptations

A number of specific adaptations were noted in the cases above using the exploitation – exploration framework to underscore the usefulness of this framing. Some of the adaptations were successful, others led to problems – successful adaptations were learnt and have been used again in less demanding situations, while unsuccessful adaptations have been dropped from the repertoire. In this section I briefly review adaptations in the "free fall" case, using Miller's framework of responses to information overload (Miller, 1960) as a general guide, but adapting it to illustrate the exploration – exploitation dialectic.

Miller proposed a scheme for describing the adjustments a system makes to information overload; it was specifically intended to apply widely across hierarchical levels, from the cell to the organ, the individual, the local work group, and ultimately the larger social institution. His scheme has 7 categories, but the ordinal relation among them is not defined – *ie*, they are not a hierarchy. Briefly, his original categories were:

- Omission – temporary non-processing of information (or new work demands)
- Error – processing work incorrectly now in the expectation of being able to return to normal processing later
- Queuing – delaying response during a period of high input in the expectation of catching up during a lull
- Filtering – attending to some categories of input while neglecting others
- Cutting categories of discrimination – responding to the input in a general way but with less precision than would be the case under lower demand
- Parallel or decentralized processing – processing inputs through two or more channels at the same time
- Escape – abandoning the task altogether

In this explication, some of Miller's categories might appear as either exploitation or exploration; it is how a strategy is used and in particular the familiarity with its specific application in a setting that determines where it should fall into the exploration or exploitation class.

Sutcliffe and Weick have built on this understanding of overload and applied it specifically to the area of organizational performance (Sutcliffe & Weick, 2008); their formulation goes beyond a simple "information – processing" paradigm (think of the famous television episode of Lucy and the candy factory (Job Switching, 1952)). It conceives of overload as not necessarily a case of too much data (or information, or work) but rather focuses more on the difficulties of making sense out of demand, capabilities and context. Other studies of critical events related to overload of some kind have taken a similar stance (Snook, 2000); that the task of actors in the system is to "continually make sense of an unexpected and dynamic situation that is characterized by unfamiliarity, scale, and speed of escalation" (Flin, 1996).

3.2.1 Adaptations of exploitation

The most frequently used adaptation in the ED is queuing. That patients queue for resources (beds, laboratory or radiology tests, the attention of doctors or nurses) is so commonplace in EDs it is considered the state of nature in that world. The existence of triage systems, in which patients are sorted into priority order based on a brief

assessment (typically of their complaint, general appearance, and vital signs) is a formalization of the queuing strategy. In the free fall scenarios, “normal queuing” was simply expanded; patients who were triaged at a level requiring them to be sent directly back to a treatment area without waiting in the waiting room were still sent back – only to wait for attention in hallways, or in chairs when there were no empty stretchers. In addition, deferrable work, such as stocking or charting was also queued in the hope that it could be completed once some slackening of demand gave respite.

However, as time wore on, queuing of deferrable work ultimately turned into filtering, in that some tasks, such as stocking, simply were dropped. Another instance of filtering involved use of the chairs to self-triage, as patients who could not maintain postural tone could be clearly identified for priority attention, while attention to the others was deferred.

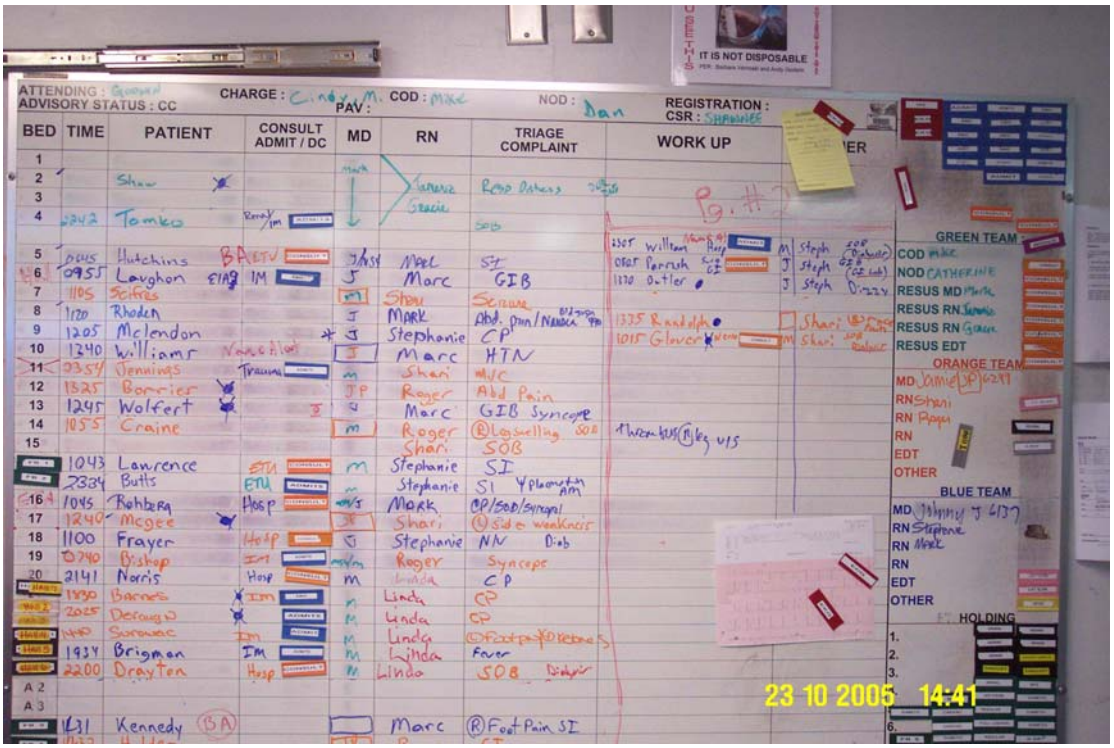


Figure 3. ED status board showing "board within a board" with reduced level of detail.

An additional adaptation was reducing precision, or cutting the level of detail in some categories. The ED status board (a user-created artefact used for distributed cognition, communication and coordination) (Wears & Perry, 2007; Wears, Perry, Salas, & Burke, 2005) provides a clear example of this strategy. Figure 3 shows an

ED status board under conditions of moderate overload. In this instance, the board had no more space for additional patients, and a "board within a board" was created. Notice particularly how the entries in this supplemental board are truncated, providing only a minimalist summary of a patient's work trajectory compared to entries made previously under more nearly normal circumstances.

All these adaptations had been used to cope with less stressful overload situations in the past; they were familiar to workers in the ED, and thus constitute a repertoire of responses to "normal, natural troubles" that can be exploited to maintain the ED's margin for maneuver.

3.2.2 Adaptations of exploration

The free fall events were unexampled, and so brought forth more novel adaptations of exploration. In one instance, the adaptation of cutting categories was used, in that hallway beds, which had previously been used only for admitted patients, were pressed into service as spaces for the evaluation of newly arrived, incompletely evaluated patients. This strategy was later abandoned due to an adverse event in one such patient, even though the causal connection between the adaptation and the patient's outcome were quite unclear.

There were two novel uses of the response of decentralization and/or parallelization. First, since physical congestion became extreme in free fall and it was difficult for staff, much less patients, to change location, staff began managing those patients whom they were near, whether they were their own patients or not. Second, the scope of authority for physicians in training was expanded, in that interns (1st year post-graduate trainees) were allowed to make dispositions (admit or discharge) on "simple cases" unstaffed by a senior resident or attending, contrary to what had been standard practice.

Finally, omission was used as a novel strategy. This was the stopping strategy – temporary non-processing of new work, so that organizational resources (in this case, an overview of the situation, coupled with sense-making and prioritization) could be replenished.

One adaptation mentioned by Miller (escape) was not used. Although the stopping strategy might be considered an instance of escape, it was not since it was envisioned as temporary (and thus is better labeled an instance of omission).

3.3 Stopping the system – a special case of exploration

Because the strategy of stopping seemed to play an important role in the free fall case, stopping will be explored in greater detail in the model. Stopping seems a special case of the exploration class of reconfiguration strategies[11]. It differs from other reconfigurations in that it does not attempt to maintain work output during the reconfiguration, but assumes (for the most part) that the work system will be restarted at some later time and in a better state. Thus it is not necessarily a strategy of surrender (although in the free fall episode it certainly felt that way), but rather an extreme case of deferring work until a later, more opportune time (although instead of deferring less essential work, all work is deferred). In this section I explore what types of stopping strategies there are, and how they might benefit (or degrade) the system's performance in achieving its goals. Specifically, stopping strategies can be invoked at the input, process, and output levels of a work system; the specific technical context determines where a stopping strategy is most usefully applied.

Stopping strategies seem more common than one might think. The clearest and most formalized example of stopping occurs in financial systems, where to avoid uncontrollable amplification from positive feedback loops, trading systems are designed with 'short circuit' mechanisms that halt trading if certain triggers (based on volume of transactions, speed of transactions (*ie*, volume per unit time), or price changes). Such measures have usually been put in place following dramatic events, such as following the "flash crash" of 6 May 2010 (Goldfarb, 2010). Essentially, these measures stop the execution of trades and indirectly the inflow of work to be done; thus they decrease the coupling between the outside world (especially high frequency trading systems) and the system of work.

---

[11] In this section I consider only stopping as a 'special case' strategy, typically invoked in the extreme. Many systems routinely stop on scheduled bases, either due to external drivers or to perform routine maintenance or refitting; these sorts of stops would fall in the exploitation class of adaptations and are not further considered here.

Another example of stopping related to the inflow of new work is the practice of ambulance diversion that for a time was common in US hospitals dealing with crowding. This strategy allowed hospitals with crowded EDs to divert incoming ambulance patients to other (generally unspecified) EDs. Although it is still practiced, diversion is falling from favour for a number of reasons. First, even though ambulance patients typically create higher workloads for staff than "walk-in" patients, their contribution to the total workload is not high (typically around 10%). Second, because crowding is a general phenomenon, diverting patients from one over-crowded ED to another (crowded, but not over-crowded) often only serves to cause the second hospital to cross a threshold and become overcrowded as well. Finally, there are substantial human costs associated with diversion – medical records and personal physicians are now unavailable, the travel burden on family and friends increases, and confusion about where a patient is becomes a problem for both patients and caregivers.

A somewhat different form of stopping as a rescue strategy is seen in nuclear power plants, where one option for controlling an emergency situation is to stop the nuclear reaction (and thus power generation); this is called "scramming" the reactor. Here instead of stopping input, the strategy stops the work process and its output (although there is a time lag between the decision to stop, and the actual cessation of the physical process).

A further example of the stopping strategy involves stopping at the output level. For example, Roe and Schulman describes use of rolling blackouts – a form of partial stopping – as the ultimate management strategy open to operators controlling the electrical power transmission grid during the 2000-2001 California energy crisis (Roe & Schulman, 2008; Schulman & Roe, 2007). Here, system integrity and survival was supported by temporarily limiting the work output.

There are many uncertainties about stopping as a resilient strategy – the phrase itself seems contradictory. A key issue is when to invoke stopping – under what conditions is it essential, *vs* potentially beneficial but optional compared with other possible strategies, *vs* harmful? Stopping too soon, or stopping unnecessarily creates costs in terms of lost output, and may entail additional costs in restarting or repairing the

system from the consequences of sudden stopping; but stopping too late, as has been broached in the free fall case, brings the risk that the window of opportunity to regain control may have passed (Dekker & Woods, 1999).

A related question is how long a stopping strategy should be maintained. How might operators know when it is safe to restart, or advisable to restart? What issues and risks arise from restarting in itself? It is interesting to note in this regard that a number of celebrated accidents (*eg*, Chernobyl, or Hurricane Katrina) occurred after the excursion into an unexampled space had ended and controllers tried to return to normal operations.

Temporary stopping has often been recommended in crisis response, to enable re-examination of the situation or facilitate engagement of additional resources and thinking on the problem (Argyris & Schön, 1974). But, Rudolph and Repenning have proposed distinguishing crises of novelty from crises of volume; crises of novelty are incomprehensible problems, those for which the system has no ready response, while crises of volume are a series of problems for which well-honed responses are available, but in which the number of problems exceeds the time or resources available to deal with them (Rudolph & Repenning, 2002). They argue that in a crisis of volume, people often do not recognize the impending disaster until it is too late so that a strategy of stopping to collect one's wits (so to speak) might be precisely the wrong thing to do, since the window of opportunity for effective action might then be missed (Dekker & Woods, 1999), or, if the system is near a "tipping point", a delay might cause it to cross that threshold and so become trapped in a vicious cycle of accumulating work and declining performance.

A final question is what, exactly, is gained by stopping? It affords an opportunity to regain, reinforce, or rebuild the system's margin for maneuver by allowing resources to be diverted temporarily from ordinary work to this sort of capacity-restoring work, and thus to potentially return the system to a more stable, more resilient, state. If stopping were used in this way, it might conceivably have value in a crisis of volume, but the circumstances under which that might or might not obtain are not entirely clarified.

French summary of Chapter 3

Ce chapitre se sert de l'abstraction pour résumer les adaptations indiquées dans les quatre études de cas présentées précédemment, afin de les placer dans deux grandes catégories, à savoir l'exploitation et l'exploration (Cuvelier & Falzon, 2011;Fu, 2007; Lengnick-Hall & Beck, 2009 ; James G Mars, 1991; Maruyama, 1963; Pariès, 2011; Piaget, 1967). Cette classification n'est destinée à être ni sans ambiguïté ni, mais plutôt utile au développement d'une représentation de la dynamique de la résilience.

L'exploitation comprend un ensemble de réponses généralement familières à des problèmes également familiers. Cette exploitation se sert d'une sorte de dynamique interne en ce qu'elle permet au système d'absorber certains chocs, tout en continuant de fonctionner à un certain niveau dans son état actuel. L'exploration, par contre, consiste fondamentalement de reconfigurations importantes et originelles du système, de changements de fond qui permettent de continuer à fonctionner, tout en se transformant à un nouvel état. En plus de permettre l'exploitation, face à des défis internes ou externes, elle soutient un apprentissage organisationnel par des moyens d'expérimenter de nouvelles manières de fonctionnement.

Ce qui unit ces deux régimes, au-delà des correspondances évidentes dans leurs parties, est la notion que bien que les deux types de réponses existent dans une sorte de tension dynamique, les deux sont nécessaires à la réussite à long terme (Carlson & Doyle, 2000, 2002; Fu, 2007), l'un doit être équilibrée contre l'autre dans un compromis fondamental (Hoffman & Woods, 2011). Un système trop dépendant de l'exploitation - trop bien adapté à son environnement actuel - serait incapable de se reconfigurer et changer en réponse à un changement de circonstances. Inversement, un système trop dépendant de l'exploration serait trop inefficace pour prospérer. Ceci constitue une occasion optimale - compromis de résilience (Woods & Branlat, 2011B): pour réussir dans le long terme, les systèmes complexes doivent toujours être partiellement «déréglé», de sorte qu'en cas d'insuffisances (assez petites) dans les réponses d'exploitation conduise à des réponses d'exploration, ce qui pourrait mener à des comportements tout neufs (Lundberg & Johannson, 2006). Un objectif particulier de ce travail est d'examiner la possibilité d'identifier lesquels des facteurs synergiques

ou antagonistes qui influeraient le type d'adaptation qui dominer, et les indicateurs de changements imminents (souhaitables ou non) entre l'exploitation et l'exploration.

La section suivante résume brièvement les adaptations déjà notées dans ces deux grandes catégories. Par exemple, les adaptations de l'exploitation utilisées couramment dans le SAU et aussi en «chute libre» y compris files d'attente, de diminuer le niveau de précision ou de détail dans certains travaux, ou de différer non essentiels du travail (comme le stockage, ou la cartographie). Adaptations de l'exploration ont été utilisés largement dans des situations extraordinaires, et inclues la décentralisation (desserrant le degré de supervision requis de personnel de niveau junior), et omission - la stratégie d'arrêt, non-traitement de nouveaux travaux afin que les ressources organisationnelles puissent être restaurés.

Le chapitre se conclut par une discussion plus détaillée de l'arrêt comme une stratégie adaptative. L'arrêt peut prendre plusieurs formes. Dans les systèmes financiers, on peut arrêter le système en cas de crise en arrêtant l'afflux de nouvelles commandes pour les métiers ; de tels systèmes ont l'avantage considérable que les travaux différés ne s'accumulent pas à être traités plus tard, mais peuvent être simplement ignorés. Les systèmes tels que les services d'urgence ont peu de capacité pour arrêter l'afflux. Cette même capacité est réduite par le problème que tout travail reporté ne fera que s'accumuler pour être manipulé tard, et certains patients peuvent se détériorer lors de l'arrêt et auront besoin d'encore plus de travail au moment où on les traitera.

Il est également possible d'arrêter au niveau du procédé, par exemple, le «scramming» d'un réacteur nucléaire n'arrête pas l'entrée des données, mais arrête le processus (bien que soumis à un certain temps de latence). Enfin, il est également possible de l'arrêter au moment de la sortie. Dans l'étude de cas de distribution électrique,les pannes roulantes - une forme d'arrêt partiel - ont été utilisées comme stratégie de gestion ultime pour les opérateurs lors de la crise d'énergie californienne de 2000-2001 (Roe & Schulman, 2008; Schulman & Roe, 2007). Ici, l'intégrité du système et de la survie ont été soutenues en limitant temporairement la sortie du travail.

Il y a beaucoup d'incertitudes sur l'arrêt comme une stratégie de résilience (la phrase elle-même semble contradictoire!) Les questions clés comprennent quand s'arrêter,

pour combien de temps, ou si d'invoquer l'arrêt complet ou partiel. L'arrêt soulève également des questions sur les dangers de redémarrage (Tchernobyl montre que ces dangers ne sont pas négligeables). Enfin, bien que l'arrêt de regrouper est souvent recommandée comme une stratégie de gestion de crise, il ya certains pensaient que les crises de la nouveauté et les crises de volume doivent être distingués. Dans les crises de la nouveauté, l'arrêt peut se permettre du temps pour chercher ce qui est logique afin de mieux comprendre la situation et élaborer des cours de l'action, mais dans les crises du volume, l'arrêt pourrait en fait aggraver la situation, de sorte que le système puisse traverser un «point de basculement» et devenir incontrôlable.

Chapter 4. Dynamics

Although the state space and stress-strain models used above have been useful in understanding these events, and particularly in providing a language in which they can be more meaningfully articulated, they both share the problem of being static representations of dynamic events[12]. Both acknowledge some sorts of transformations – state transitions or deformations – but explanations of how these transformations come about, how quickly, what promotes or retards them, *etc*, are missing.

And, dynamic events pose significantly greater problems for understanding, in that they involve not only predictions about what a system will do in response to an event, but also how that reaction will "… in turn influence the future development of the process and the interaction" (Hollnagel, 1993).

Nemeth has pointed out that traditional system representations have been developed to operate in simpler and more static circumstances that can be readily represented by simple, static diagrams, but that "… resilience is substantially about dynamic, not static, properties. New thinking along the lines of resilience requires new kinds of language to describe system properties" (Nemeth, 2009). Dijkstra has noted that "... our intellectual powers are rather geared to master static relations and that our powers to visualize processes evolving in time are relatively poorly developed" (Dijkstra, 2008). Similarly, Dekker has noted the need for new models of accidents (and by extension, performance under constraints, ambiguity, uncertainty, risk, time pressure, *etc*) that are "not constituted of parts and their interactions, but as a web of dynamic, evolving relationships and transactions" (Dekker, 2005). He notes further that a common cause of accidents in complex, sociotechnical systems is the slow, incremental movement of operations towards the boundary of failure (Cook & Rasmussen, 2005; Rasmussen, 1997; Woods & Sarter, 2000), called "organizational

[12] Richard Cook is, I think, the first to specifically emphasize the time dynamics of resilience, in oral remarks at the 2nd Resilience Engineering Symposium (Cook, 2006). However, it seems implicit in many other discussions of resilience as well, such as the functional resonance accident model (FRAM) (Hollnagel, 2004).

drift" (Snook, 2000), which clearly seems to have been a factor in these "free fall" episodes. This drift is difficult to capture in the structuralist, mechanistic language (deriving from Cartesian / Newtonian reductionism) in common use in many hazardous industries, and which overwhelmingly dominates discourse in healthcare (Dekker, 2011; Hunte, 2010). The fundamental goal of this thesis is to extend our previous models of resilience in ways that support insight into the dynamics of resilience, in at least descriptive and ultimately prescriptive ways; to develop Nemeth's "new language" in support of Dekker's "new models".

Events, behavior and structure are often presented as a hierarchy of ways of looking at the world with events arising from underlying behavior patterns and behavior arising because of structure. Simulation models represent the structure that generates the behavior; we see and understand this link between structure and behaviour through simulating the system. The connection between structure and behavior is strong, but it can be difficult to gain an understanding of how structure causes behavior. System dynamics models using computer simulation enable one to quickly gain insight into this connection. It is important to emphasize here that we are using the term 'structure' not to refer to formal organizations of work (such as might be represented in a hierarchical, organization chart) but rather to refer to the relationships among causal influences and feedback loops present in the work system. It is in this sense that "structure drives behaviour."

Although the initial motivation for the models was an overcrowding crisis in a medical setting, they are intended to apply more generally to a broad range of situations involving overload conditions or novel challenges and threats in complex sociotechnical systems. These situations are characterized by 1) action-based inquiry, in which interpretation (sensemaking) and action (choice) are closely linked; 2) temporal dynamism, in that the situation changes on its own even if no action is taken; and 3) endogeneity, in that actions change the nature of the problem and the problem-solving environment, so that understandings and actions co-evolve and are reciprocally determined (Rudolph, Morrison, & Carroll, 2009).

French summary of Chapter 4

Ce chapitre est une brève argumentation pour le développement de modèles dynamiques de la résilience. Tant d'espace d'état et de contrainte-déformation des modèles précédemment sont des représentations statiques d'événements dynamiques. Mais les événements pose dynamique nettement supérieure à la compréhension des problèmes, en ce qu'elles impliquent non seulement des prédictions sur ce qu'est un système fera en réponse à un événement, mais aussi comment cette réaction «... à leur tour influencer le développement futur du processus et l'interaction» (Hollnagel, 1993).

Plusieurs chercheurs ont fait écho à cet appel pour plus de représentations dynamiques de la résilience. Notes Nemeth que «... la résistance est sensiblement sujet dynamique et non statique, les propriétés. Une nouvelle réflexion sur le modèle de la résilience exige de nouvelles formes de langage pour décrire les propriétés du système» (Nemeth, 2009). Dijkstra a soutenu que «... nos facultés intellectuelles sont plutôt orientées à maîtriser les relations statiques et que nos pouvoirs de visualiser les processus qui changent dans le temps sont relativement peu développés »(Dijkstra, 2008). De même, Dekker a noté la nécessité de nouveaux modèles d'accidents (et par extension, la performance sous contraintes, l'ambiguïté, l'incertitude, de risque, la pression du temps, et cetera) qui «ne sont pas constitués de pièces et de leurs interactions, mais comme une toile de dynamique, l'évolution des relations et des transactions »(Dekker, 2005).

Le mouvement lent et incrémental des opérations vers la frontière de l'échec (Cook & Rasmussen, 2005; Rasmussen, 1997; Woods & Sarter, 2000) (Snook, 2000), semble avoir été un facteur dans ces «chute libre» épisodes. Cette dérive est difficile à saisir dans le structuraliste, le langage mécaniste (dérivant de cartésiennes / newtonienne réductionnisme) d'usage courant dans de nombreuses industries dangereuses, et qui domine massivement le discours de la santé (Dekker, 2011; Hunte, 2010). L'objectif fondamental de cette thèse est d'étendre nos modèles précédents de la résilience de manière à soutenir un aperçu de la dynamique de la résilience, dans au moins des

moyens descriptifs et finalement prescriptifs; à développer Nemeth «nouveau langage» à l'appui de Dekker «nouveaux modèles».

Chapter 5. Building a model

The model-building work proceeds from the specific events of 'free fall, using them to build a theory of the dynamics of resilient performance in an abstract sociotechnical system, and finally returns to ground the models and theories developed in empirical observations in the original setting; *ie*, from the specific to the general, then back to the specific.

In the exposition of the models, I do not use the more traditional separation between methods (the models' structures) and results (their performance), for two reasons. First, model-building is an iterative process, where the builder alternates between models and results as part of model development. Thus, this reflects more accurately the thinking and process of construction than would the description of a final model, followed by a detailed exposition of how its behaviours conform with reality. Second, it seems much easier to understand a model by a process of layering – starting with a very simple structure and understanding its behaviour (its physics, if you will) – and then adding progressively more complexity, reviewing the new behaviours at each additional step. The next two subsections cover this imbrication of layers, alternating between descriptions of structure and descriptions of the behaviours produced by that structure[13].

When viewing the graphical representations of the model that follow, it is important to keep in mind that these are representations of the structure of functional relations, not of components or agents (see Section 1.3). Thus the information in the Figures is contained in the arrows representing influence relations more than in the components (the accumulations, flows, and variables).

---

[13] In the model development that follows, it is often necessary to provide specific numeric values for many of the variables. These have been approximated using data from the ED work system. I used Little's Law (Little, 1961) to estimate equilibrium time to completion from the known arrival rate and average number in the system for busy, afternoon and evening periods, and substantiated that estimate in discussion with domain experts. However, it is the qualitative behaviours of the system, not the quantitative results that are important to the objectives of this work. To support that, I refer to quantitative results only when needed to emphasize or illustrate a point about the qualitative behaviours.

5.1 Work activity

The diagrams in this and the following sections use common conventions for representing system dynamics models. Variables in boxes (such as *Work in Progress* in Figure 4) are stocks, which are accumulations or reservoirs, like water in a bathtub. Variables show by double arrows and 'valves' (such as *New Work Rate*) are flows, which act to increase or decrease stocks. Variables influencing stocks, flows, or other variables (like *time to completion*) are shown in lower case. Arrows show these influences: a '+' sign near an arrowhead indicates that changes in two variables move in the same direction (*eg*, if *Work in Progress* increases over its current level, then *mismatch* also increases over what it would have been absent a change in *Work in Progress*; a '-' sign indicates they move in opposite directions.

5.1.1 Structure

The simplest, basic structure is the work subsystem shown in Figure 4. Work demand comes from outside the system at some rate represented by the variable *New Work*. It accumulates in the stock variable *Work in Progress*, and is completed at a rate defined by the *Completed Work* variable[14]. *Completed Work* depends on both the volume of pending work (*ie*, the magnitude of *Work in Progress*), and the *time to completion*. In turn, *time to completion* depends on a set of influences: the intrinsic nature of the task, represented by *av time to completion*, and a level of *mismatch* in the system, the mismatch between demand (*Work in Progress*) and *capacity*.

[14] Although it is often easier to speak of work as discrete items being processed (*eg*, widgets, or patients), I do not assume discrete entities are involved here, for in some systems the inputs, outputs, or both might be continuous. For convenience here I will refer to them as widgets, and use them to represent some unit of the work effort required. This affords some representation of the differing complexity of problems that present; *eg*, in an ED, complex or critical cases require more work effort (more widgets) than simpler ones. Similarly, the units of time are arbitrary, but for convenience I will use hours.

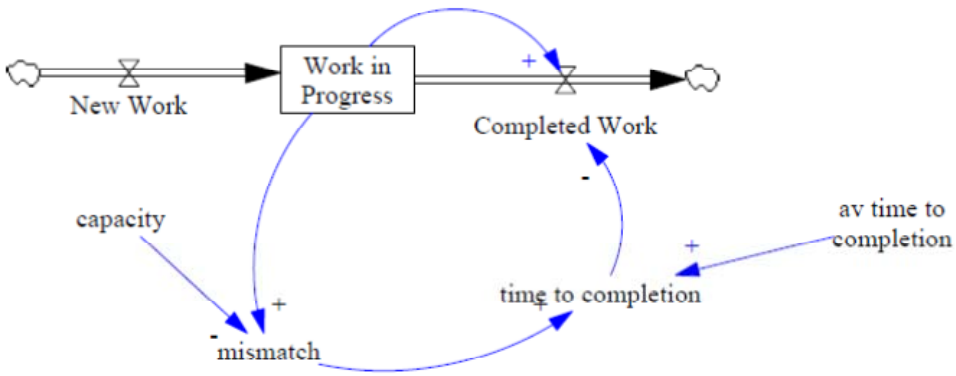


Figure 4. The basic work system.

This simple model contains one causal loop (Sterman, 2000). As *Work in Progress* increases, the *mismatch* between *capacity* and demand also increases. (By using *mismatch* as a mediator of overload, the model is treats both increases in demand and decreases in capacity as equivalent in their effects). This initially causes an improvement in performance (an arousal-like effect), but with increasing mismatch eventually shifts to degraded performance, *ie*, increase in *time to completion*, which then decreases the rate at which work is completed. This nuance is not strictly necessary, as for the most part the explorations will take place in areas far beyond any arousal effect, but was included because many stakeholders in the ED world mentioned it (*eg*, taking pride in their ability to "rise to the occasion"), and because there is at least some thinking that this property is a reasonable description of performance response in a variety of settings (Kahneman, 1973; Zijlstra, Roe, Leonova, & Krediet, 1999). Please note that in this model, the means by which higher values of mismatch cause degraded performance are not further specified; they can be a combination of congestion or turbulent flow, requirements for greater coordination work, errors or shortcuts in work requiring additional re-work, queuing for critical resources, *etc*. A decrease in outflow (*Completed Work*) then tends to increase *Work in Progress* even further. Thus, this loop, which represents overload, is a positive feedback loop. Positive feedback is amplifying, and thus potentially destabilizing; if unchecked, it will cause a system to expand exponentially, and thus spiral out of control.

Without loss of generality, I make three simplifying assumptions in this abstract model. First, I leave out the case where some variables other than the rate of completed work and volume of pending work are the key state variables which we wish to control; for example, in financial systems, price might be such a variable. Second, for the purpose of this model I consider control to mean maintaining the state variables within certain reasonable bounds, and/or returning them rapidly to those bounds if they are exceeded. Again for simplicity I leave out the case where control might be interpreted as stability, the avoidance of rapid fluctuations in value rather than keeping to a preferred range of values. Third, I treat throughput as the summary indicator of system performance. In real systems, of course, operators might trade off other properties, such as work quality, instead of volume in order to maintain control, but for simplicity we will consider only a single performance variable. The fundamental argument for these simplifications is that if we see complex and interesting behaviours in an extremely simplified system, we should also expect those complexities to appear in more complex and more nearly realistic systems.

For simplicity of exposition, challenges to the system are presented here only as increased work demands, as overload[15]. Overload is both a common and important stress on many different sorts of systems, and has been clearly implicated in the catastrophic collapses of complex systems in multiple domains (Rudolph & Repenning, 2002; Shanker & Richtel, 2011; Weick, 1990). In addition, overload (produced by a combination of increased input (demand) and slowed work output due to decreased capacity) played a prominent role in the 'free fall' episodes that motivated this work. Thus, studying a system's response to an overload challenge can provide insight into its resilient abilities and potentially can shed some light on the phenomenon of 'free fall.'

The effect of overload (demand / capacity mismatch) on throughput was estimated from ED data and validated in discussions with ED staff. It is expressed as a multiplicative factor, which is near one for levels of crowding below the system's

[15] Clearly there are many other types of threat to system performance (*eg*, degradation of capacity, loss of adaptive ability, slowdowns in processing, *etc*. These scenarios have also been explored but are not presented here as they did not result in qualitatively different behaviours.

functional capacity, and which increases at an increasing rate as the system becomes progressively more congested, until a ceiling value is approached. This sigmoid functional shape leads to model responses that are qualitatively similar to those expressed by workers in the ED domain: that the system is able to cope effectively up to a point (think of the elastic region), then gradually starts to fall behind (think of the plastic region), and then quite abruptly, slows dramatically (think of a deformation)[16].

5.1.2 Behaviour

The following graphs illustrate the 'physics' of this simple system[17]. They are useful baseline behaviours to be contrasted against the more complex models to follow, which include representations of exploiting and exploring activities.

We first consider a "null state" in which overload has no effect (by setting the effect of *mismatch* on *time to completion* to one). Figure 5 shows this system's response to a pulse challenge, starting in a steady state (*ie*, in static equilibrium), and assuming that overload has no effect, *ie*, that the system is perfectly resilient in that its performance can rise to meet any level of demand[18]. At hour 60, the *New Work* doubles for a period of 4 hours and then returns to normal (blue line). *Work in Progress* (green line) and *Completed Work* (red line) then begin to rise, reaching peaks at hour 64 and returning to baseline by hour 69, 5 hours after the pulse ended.

Figure 6 shows that this perfectly resilient work system can respond effectively to pulse challenges of increasing magnitude (specifically, 100%, 200%, and 300% increases in demand). The stock *Work in Progress* increases transiently in each challenge because of the lag between the increase in arrival rate and the increase in work completion rate, but eventually the system's response catches up to match the

---

[16] In the black humour of the ED, this is commonly known as "clotting off," referring to the physiologic phenomenon of arterial thrombosis – a clinical disaster. Similar nonlinear phenomena have been noted in fluid dynamics and in traffic flow (May, 1989).

[17] Only a small number of representative graphs are shown here for the sake of simplicity. The model-building process involved a large number of exploratory test cases, varying the magnitude, duration, and shape (*eg*, pulse, single step function, ramp, periodic oscillation, *etc*). These further complexities will only be presented here where they result in substantive differences in behaviour, but are available on request.

[18] Such pulse tests are of course not very realistic, but are widely used in analysis of dynamic systems because they help provide a clear picture of how the system acts in disequilibrium situations.

demand placed on it. When the rate of new work returns to baseline, the backlog is "worked off" and *Work in Progress* also returns to its baseline value. This behaviour corresponds to the elastic zone in the stress-strain representation. It is interesting to note here that although the magnitude of *Work in Progress* increases in proportion to the size of the challenge pulse, the time needed for the system to recover to baseline after the challenge is removed is very short (1 to 1.5 hours, and increases only slowly with increases in pulse magnitude.

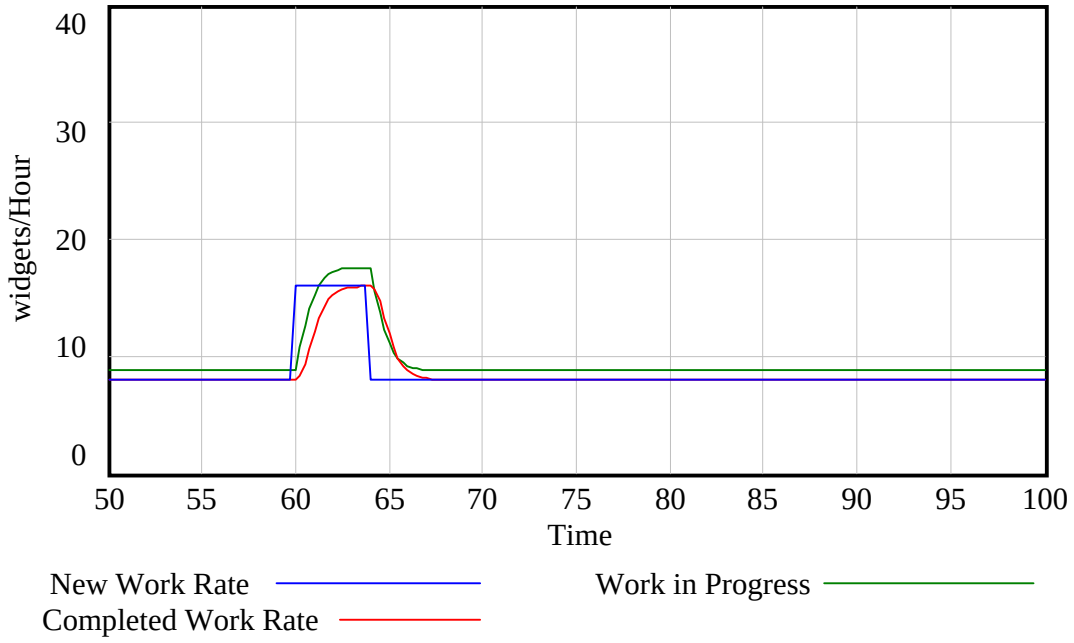


Figure 5. Response of perfectly resilient work system to pulse overload

But, no system is perfectly resilient, so a more realistic model includes an effect of overload, of the mismatch between demand and capacity as described above**.** Figure 7 shows this system's response to the same pulse demand; because the time for completion now increases due to overload, *Work in Progress* rises to a higher level, although the system is able to return to baseline within almost the same amount of time (5.5 *vs* 5.0 hours).

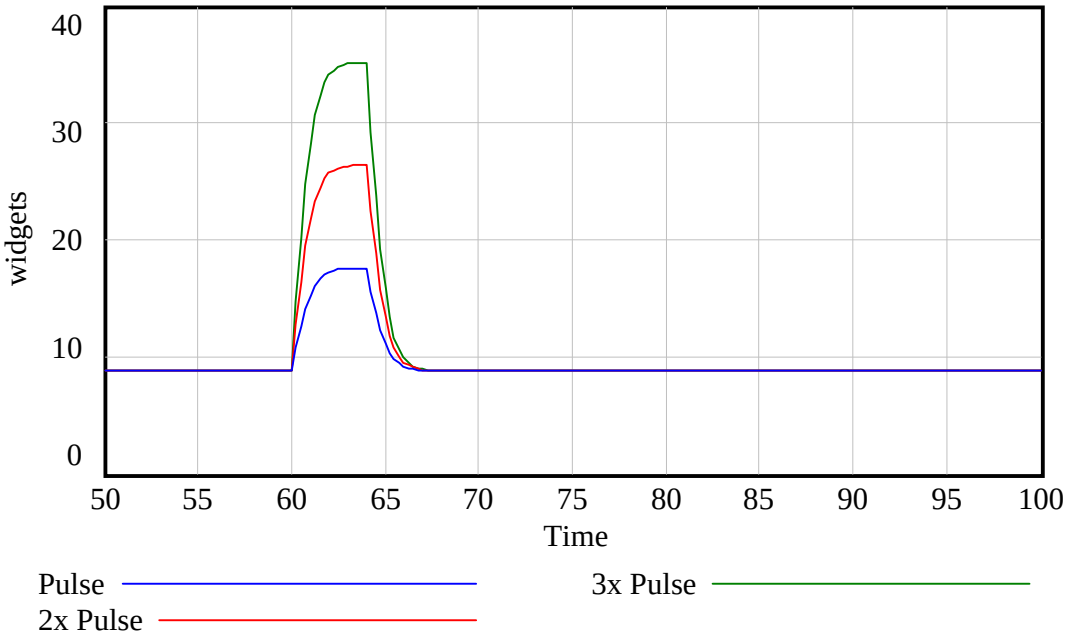


Figure 6. Response of a perfectly reslient system to increasing pulse challenges.

However, the response of this system to overload is now much more brittle. At challenges not much larger than the standard challenge above, the system completely decompensates, as shown in Figure 8, in one of the basic patterns of adaptive failure (Branlat & Woods, 2010). As before, an increase in demand causes an initial increase in the stock of work pending and in the *Completed Work* rate. But, the effect of increasing mismatch between demand and capacity eventually causes the completion rate to fall. Once it falls below the equilibrium rate of incoming work, *Work in Progress* shoots up, increasing *mismatch* even further, and the system collapses.

In contrast to the perfectly resilient model, where larger increases in pulse magnitude produced large increases in work being processed but only small increases in recovery time, in this brittle model, the reverse holds true. Minor increases in pulse magnitude produce only small changes in *Work in Progress*, but rather large increases in recovery time. This is a classic pattern of decompensation – falling behind the tempo of operations (Woods & Branlat, 2011a). In addition, it suggests that as a system approaches the boundary of failure, increases in recovery time might be better critical indicators of impending failure than the total volume of work being processed, or the

volume of new work arriving. This is a more dynamic measure – how fast the system is recovering, *vs* how well it is doing overall.

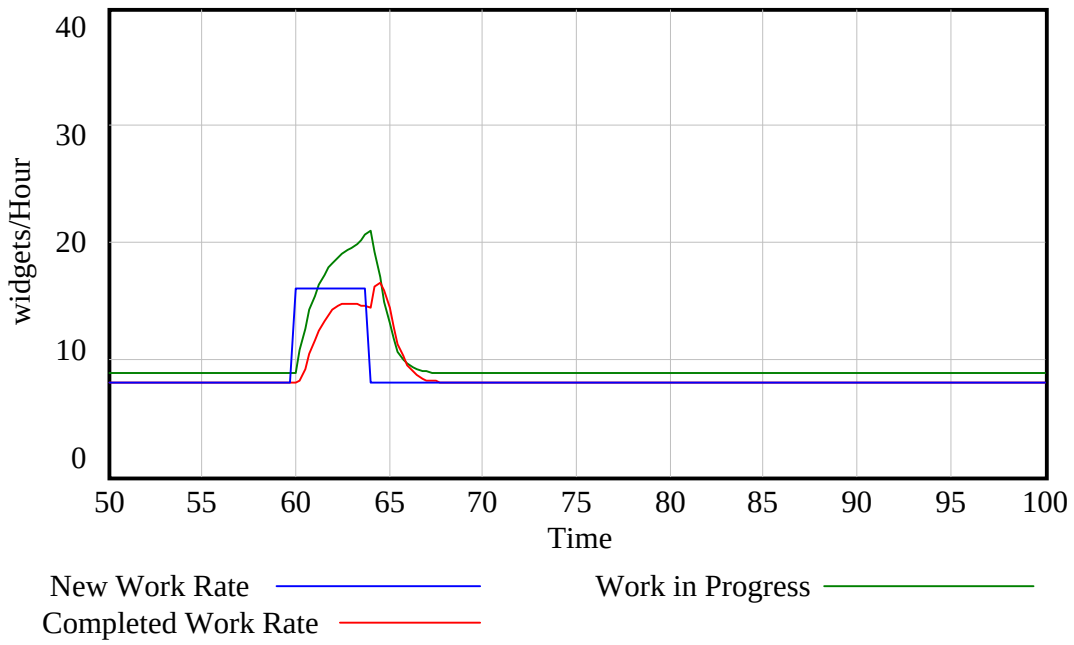


Figure 7. Response with overload degradation.

Stress-strain plots of the peak values for *Work in Progress* and of the time until recovery show dissimilar signatures (Figure 9); the impending transition from the elastic to the plastic, or 'extra' region is signaled by a dramatic increase in the length of time required to return to baseline (around pulse sizes of around 110% of baseline), while the curve for peak *Work in Progress* begins to flatten out at this point. This relatively sudden transition may be a more salient signal of approach to the boundary of catastrophic decompensation than the more gradual increases in total work in progress. A similar signature of impending decompensation has been described in natural ecological systems (Scheffer *et al.*, 2009; Woods, 2011), and has been justified on theoretical grounds.

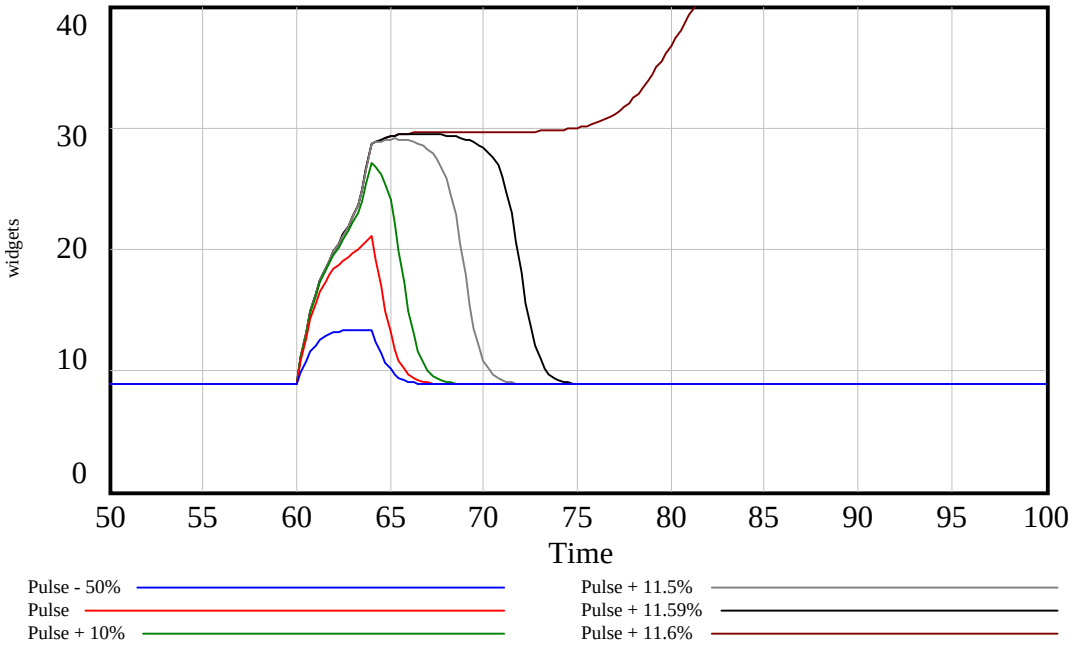


Figure 8. Brittle response to increasing overload.

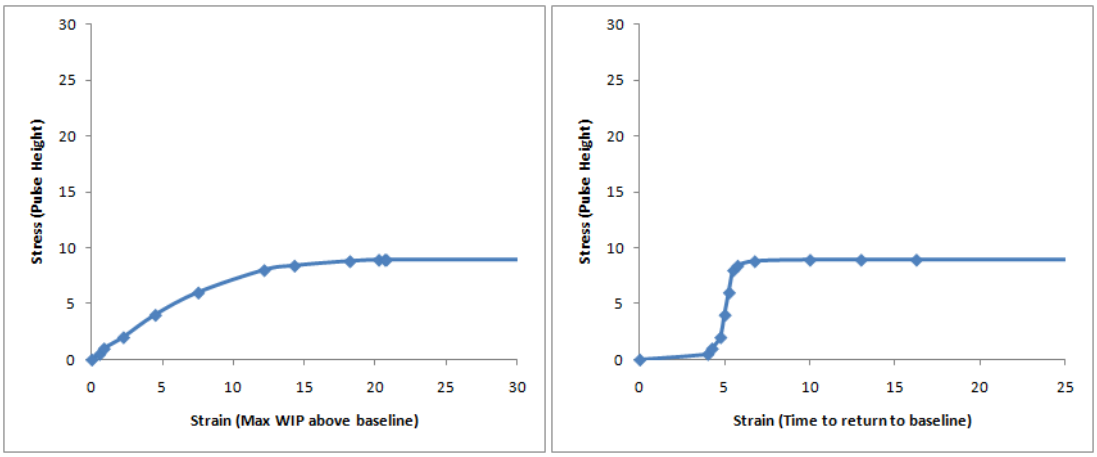


Figure 9. Stress-strain plots of decompensation.

5.2 Exploitation

Systems typically have additional resources they can exploit to protect against the sort of catastrophic decompensation noted in Figure 8. These can take the form of buffers, short-cuts or tradeoffs that constitute a "bag of tricks" operators use in order to meet their goals. In the ED such compensatory strategies might include buffering (*eg*, by

storing patients in hallways and other non-standard spaces), or deferral of some work activities (*eg*, charting, or stocking). Therefore, a more useful model adds a stock of resources that can be drawn down to compensate for a shock.

5.2.1 Structure

Figure 10 illustrates this addition; for simplicity I have represented only a single stock of resources, *Margin*. Their consumption is triggered by the mismatch between *capacity* and *Work in Progress*, and increases as *mismatch* increases; thus the work completion and adaptive resource consumption are co-flows. This consumption decreases *time to completion*, countering the effect of mismatch. Resources consumed must eventually be restored, and this restoration is slowed by *mismatch*. Both consumption and restoration of *Margin* are delayed: it takes time to perceive an overload and bring resources to bear; and it takes both time and discretionary energy to take advantage of the opportunity to restore adaptive resources previously consumed (Crossan & Apaydin, 2010).

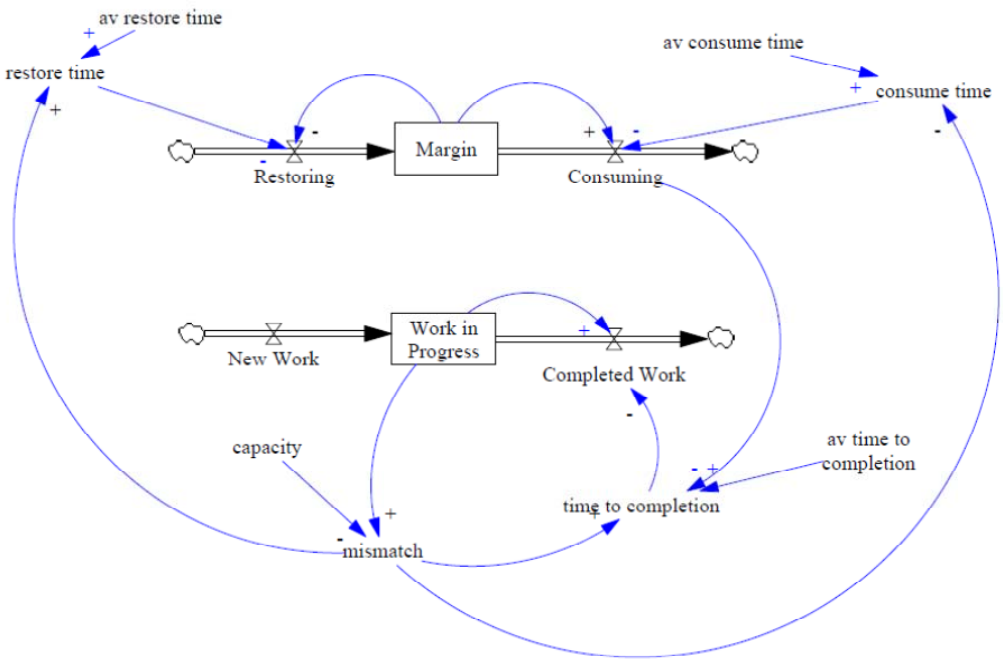


Figure 10. Exploitation of additional resources to compensate for overload.

It will be useful to discuss this more complex model in terms of the feedback loops that are embedded in it. I have already noted the positive feedback loop representing overload in the basic work system (Figure 4). For clarity, this loop is highlighted in

Figure 11; it runs from *Work in Progress* to *mismatch* to *time to completion* to *Completed Work* and back to *Work in Progress.* The addition of the notion of "margin for maneuver" creates two additional feedback loops: a negative feedback, or balancing loop representing consumption of these resources, and a second positive feedback loop representing their restoration.

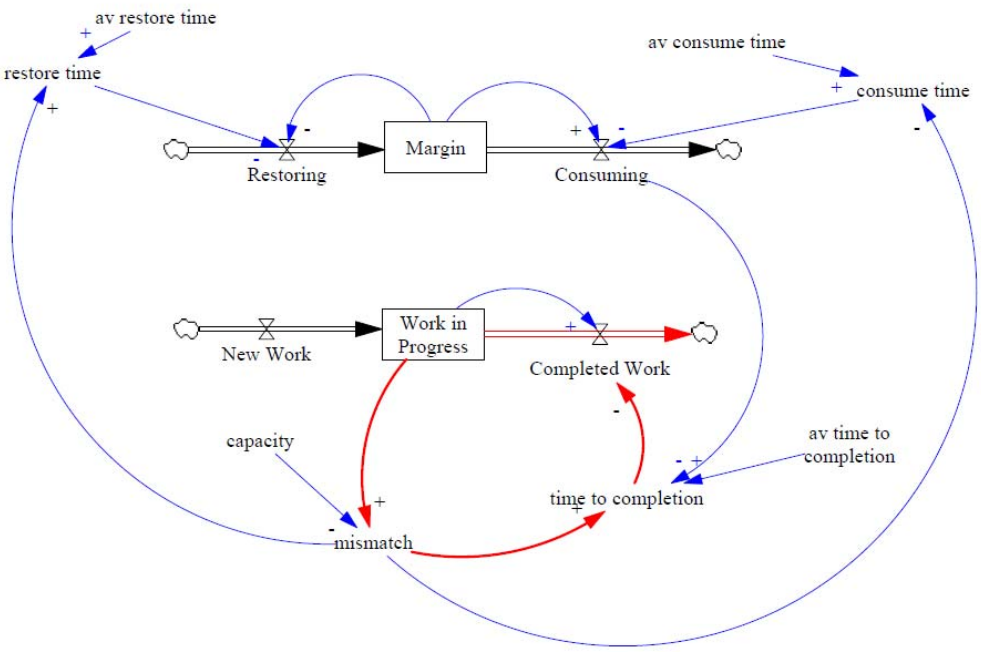


Figure 11. The overload loop; a positive feedback loop.

Figure 12 highlights the consumption loop; an increase in *Work in Progress* causes *mismatch* to increase, leading to a reduction in *consume time* and consequently an increase in the rate, *Consuming* (*ie*, margin resources are more rapidly consumed, more rapidly brought to bear). This leads to a decrease in *time to completion*, which increases the rate, *Completed Work*, thus decreasing *Work in Progress*.

Similarly, Figure 13 highlights the positive feedback (reinforcing, amplifying) restoration loop; it shows that an increase in *Work in Progress* and the consequent increase in *mismatch* increases *restore time*, which decreases the *Restoring* rate, and thus decreases the stock of resources (*Margin*). This decreases the rate at which these resources can be employed (*Consuming*) and thus ultimately results in a further increase in *Work in Progress* via increases in *time to completion* and decreases in the *Completed Work* rate.

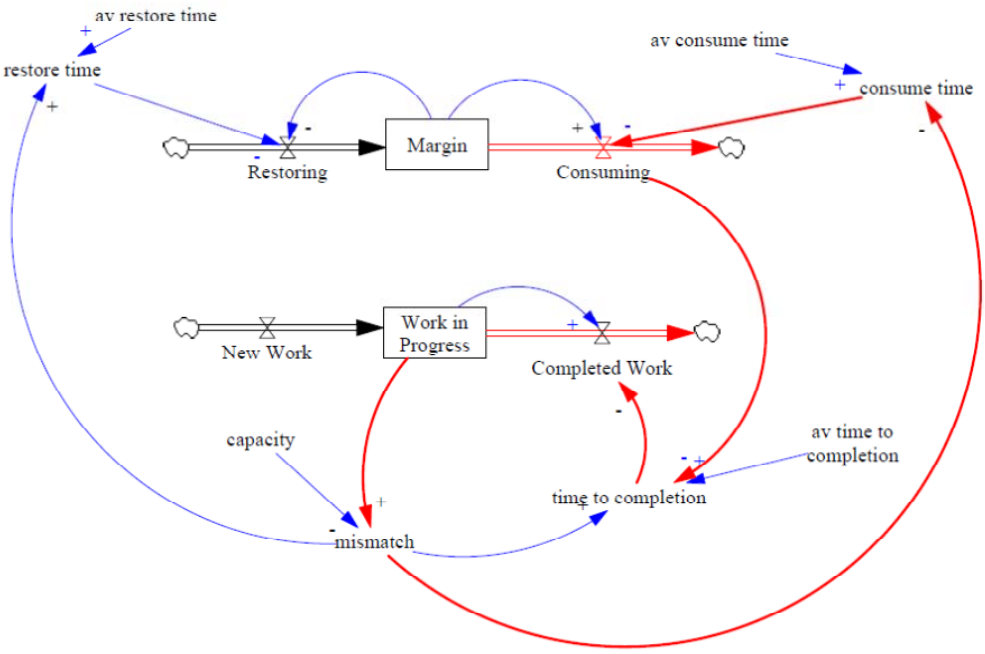


Figure 12. The consumption loop; a negative feedback loop.

The performance of this system depends on the relative balance among the positive and negative feedback loops. When the balancing loop dominates, the system will tend to be homeostatic and return to baseline conditions after a shock, while if the reinforcing loops dominate, then the system will be unstable and eventually spiral out of control. Of course, many more complex behaviours such as oscillations may also appear, depending on the time delays and relative sizes of the stocks and flows. And, one way to effect control would be by changing the relative strengths of the three loops.

5.2.2 Behaviour

The addition of adaptive resources gives the system a more resilient response to an overload shock. Figure 14 illustrates the changes in *Margin*, *Consuming*, and *Restoring* in response to a pulse challenge similar to those previous; *Margin* is consumed during the period of overload, and is restored to about its steady state level by hour 68.

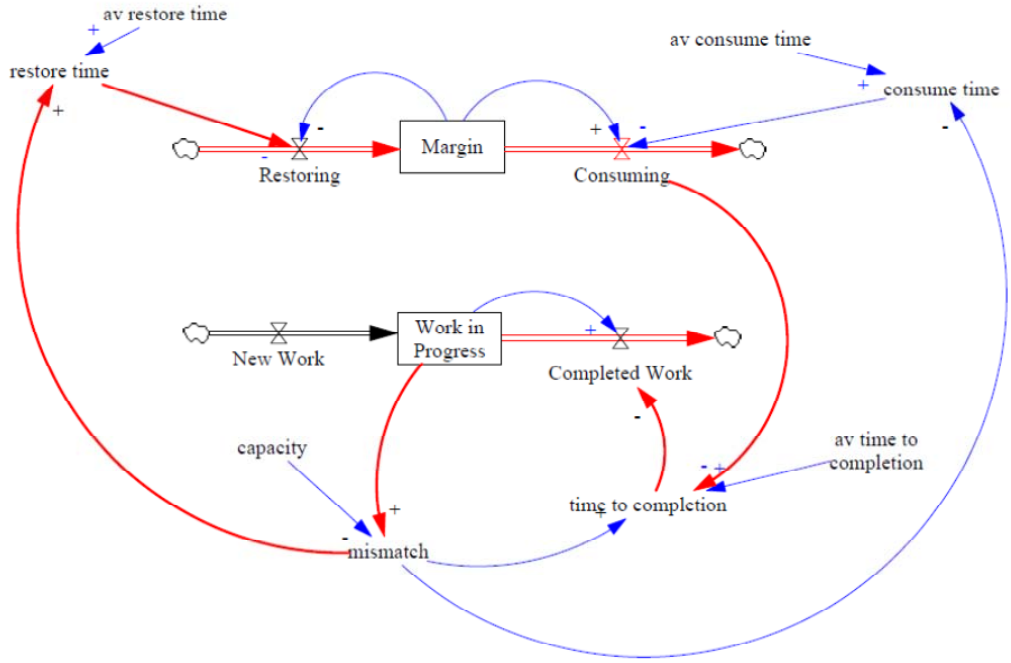


Figure 13. The restoration loop; a positive feedback loop.

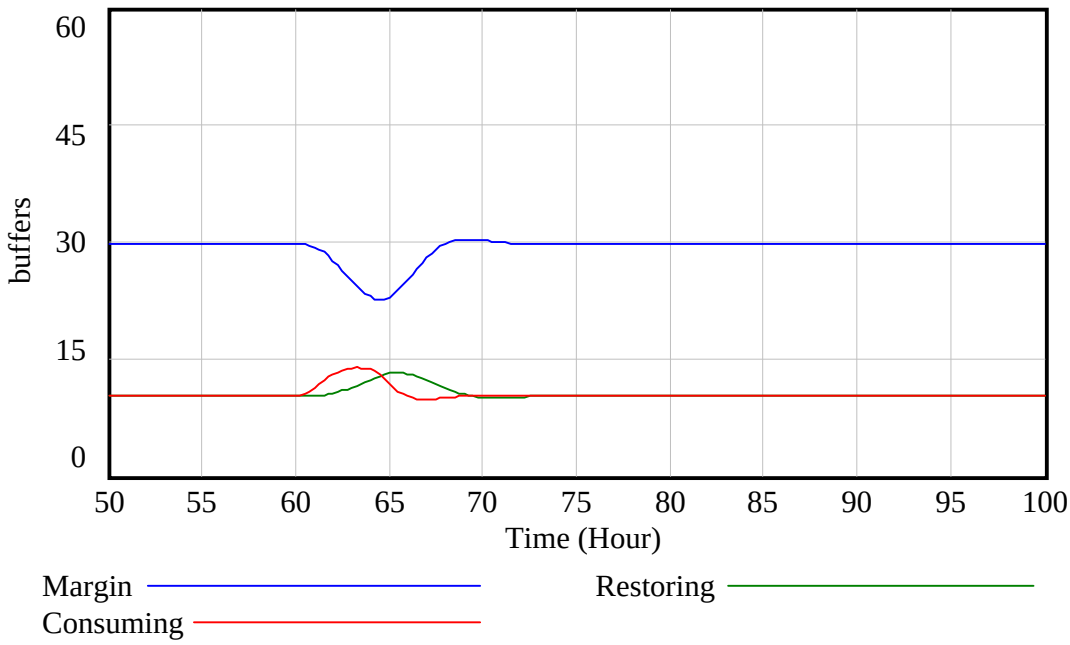


Figure 14. Response of Margin to a pulse challenge at hour 60.

Figure 15 shows the effectiveness of these resources by displaying the new system's responses (marked by a *) to increasing pulse challenges; the same pulse challenge that led to decompensation in Figure 8 is now tolerated. In addition, for somewhat

greater challenges, the system does not spiral completely out of control, but rather moves into a new steady state. This is not necessarily desirable, as the system now is "stuck" in a state of permanent overload and congestion, *even though the initiating shock has long passed.*

An important implication should be drawn from Figure 15; the system now demonstrates path dependence, in that its current state depends not only on the present value of its inputs, outputs, and current workload, but also on its history.

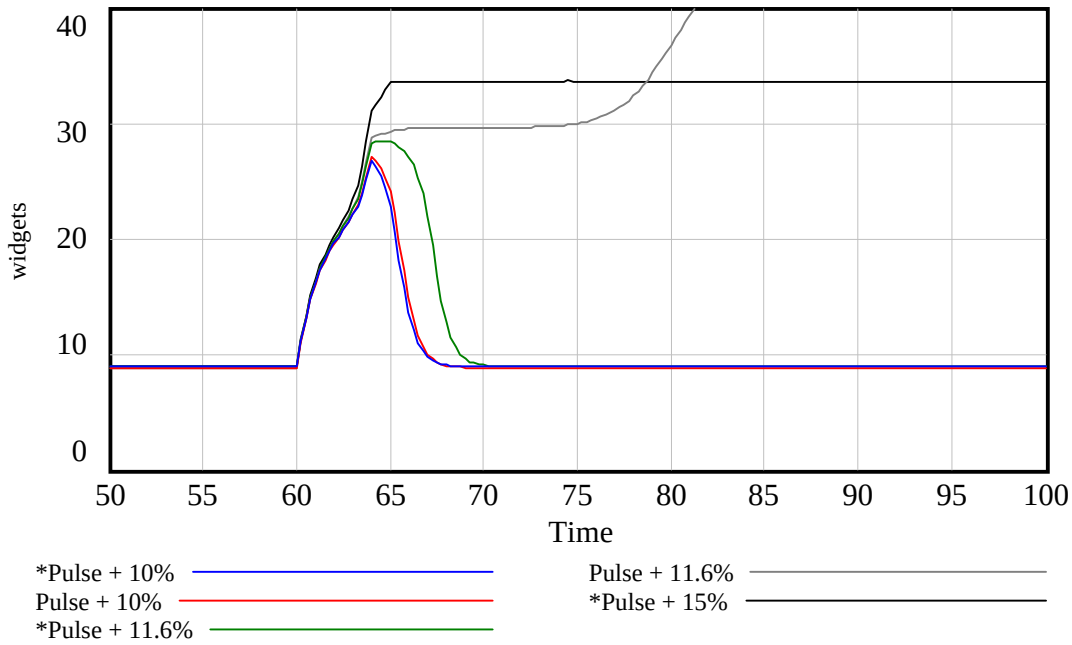


Figure 15. Adaptive resources mitigate overload shocks.

This completes the development of the system dynamics model to be used in the explorations to follow. It has shown behaviours that seem characteristic of the ED system used as a motivating example, and similarly of the disparate systems presented in the case studies. These behaviours have included:

- Elastic deformation under modest stresses, plastic deformation under more severe stress, and ultimately sudden decompensation when a threshold, or "tipping point" is crossed.
- Increasing time to recovery, or falling behind the tempo of operations, as a signature that a decompensation threshold is being approached.

- Improvements in resilient performance by strategies of exploitation, drawing on resources constituting a safety margin or margin for maneuver. These improvements mitigate the disruption caused by a challenge (*ie*, they lower the peak values for *Work in Progress*, and shorten time to recovery. By doing this they increase the threshold for decompensation.
- Path dependency and phase shifts, in that a shock of sufficient magnitude can move the system into a new state of degraded performance that persists even after the shock has passed.

French summary of Chapter 5

Ce chapitre décrit le modèle de dynamique des systèmes construits pour explorer les enjeux de cette thèse. Il n'utilise pas la séparation entre les méthodes plus traditionnelles (structures des modèles) et les résultats (leur rendement), pour deux raisons. Tout d'abord, la construction de modèles est un processus itératif, où le constructeur alterne entre les modèles et les résultats dans le cadre du développement du modèle. Ainsi, ce qui reflète plus fidèlement la pensée et les processus de construction que serait la description d'un modèle définitif, suivie par un exposé détaillé de la façon dont ses comportements conformes à la réalité. Deuxièmement, il est beaucoup plus facile de comprendre un modèle par un processus de stratification - à commencer par une structure très simple et la compréhension de son comportement, puis en ajoutant de la complexité de plus en plus, en examinant les nouveaux comportements, à chaque étape supplémentaire. Ainsi, le chapitre se poursuit par une alternance entre les descriptions de la structure et les descriptions des comportements produits par cette structure.

Le modèle utilise le formalisme du système dynamique des stocks (accumulations) et des flux (taux de changement), dont chacune peut être influencée par d'autres, variables auxiliaires. Le stock principal est appelé Work in Progress, et représente l'activité de base du système. Il est augmenté par un flux de nouveaux travaux dans le système, et diminué par un écoulement vers l'extérieur du travail accompli. Le système a une capacité, et l'inadéquation entre la demande et la capacité entraîne une diminution de la performance, qui se manifeste par un ralentissement de l'écoulement. Si discordance n'a aucun effet sur le taux de travaux achevés, alors le système est parfaitement élastique; que l'effet de décalage devient trop grand, le système affiche les modèles classiques de décompensation pour les défis de surcharge croissante.

J'ai ensuite élaboré sur ce plus simple de systèmes par l'ajout d'un stock plus, la marge a appelé, ce qui représente des ressources, des tampons et des stratégies d'exploitation utilisées pour atténuer les conditions de surcharge couramment rencontrés. L'utilisation de ces ressources les épuise, mais les compteurs de l'effet de décalage et améliore ainsi les performances. L'épuisement des ressources doit être restauré, mais

le taux de la restauration est dégradé par décalage. L'ajout de cette variable supplémentaire crée un système qui ne peut tolérer des défis qui ont causé le plus simple système de décompenser. Cependant, ce nouveau système montre également des décalages de phase - un nombre suffisamment important, des difficultés temporaires, le système pour passer à un état dégradé d'opérations, à partir de laquelle il est incapable de récupérer, même après le premier défi.

Ce modèle contient trois boucles de rétroaction, deux négatifs et un positif. La boucle de rétroaction premier résultat positif implique un travail en cours et d'inadéquation. Comme la charge sur le système augmente, augmente décalage, ce qui provoque la sortie de diminuer, ce qui provoque la charge (Work in Progress) afin d'augmenter encore davantage. La boucle de rétroaction négative implique la consommation de ressources de marge, qui tendent à stabiliser le système. Et la boucle de rétroaction positive définitive implique l'effet de décalage sur la restauration de ces ressources; la surcharge provoque une consommation accrue et une diminution de la restauration, appelée Marge d'être épuisé, ce qui accroît encore la surcharge.

Les performances de ce système dépendent de l'équilibre relatif entre les boucles de rétroaction positive et négative. Lorsque la boucle équilibre domine, le système aura tendance à être homéostatique et retour à des conditions de base après un choc, tandis que si les boucles renforçantes dominent, alors le système sera instable et finalement spirale hors de contrôle. Bien sûr, beaucoup de comportements plus complexes tels que des oscillations peuvent aussi apparaître, selon les délais et les tailles relatives des flux et des stocks.

Ce modèle très simplifié montre comportements caractéristiques du système SAU utilisé comme un exemple motivant. Ces comportements incluent:

- Une déformation élastique sous contraintes modestes, la déformation plastique sous contrainte plus sévère, et, finalement, une décompensation brutale lors d'un seuil, ou «point de basculement» est franchi.
- Augmenter le temps de récupération, ou de tomber derrière le tempo des opérations, comme une signature que le seuil de décompensation est abordé.

- Amélioration de la performance résiliente par des stratégies d'exploitation, en s'appuyant sur les ressources constituant une marge de sécurité ou de marge de manoeuvre. Ces améliorations atténuer les perturbations causées par un défi (par exemple, ils abaissent les valeurs de crête des travaux en cours, et de raccourcir les temps de récupération. En faisant cela, ils augmentent le seuil de décompensation.
- Sentier de dépendance et des changements de phase, dans ce choc d'une ampleur suffisante peut déplacer le système dans un nouvel état d'une dégradation des performances qui persiste même après le choc est passé.

Chapter 6. Exploration

In this section I use the model just developed to examine strategies of exploration via a series of stylized experiments. The choice not to embed exploration strategies in the model by extending it further was a conscious one in keeping with the minimal modeling principle (Hollnagel, *et al.*, 1995) for the following reasons. First, at this point the model is developed sufficiently to show some of the more complex behaviours seen in the case studies of real systems, and the 'free fall' cases in particular, so additional complexity is not strictly required. Second, to add these strategies to the modeling would require vastly increasing the number of assumptions that would have to be made about what sort of strategic changes might be important, what should trigger them or reverse them, what variables ought to be monitored, and so on. While I examine many of these assumptions in the experiments that follow, embedding them in the structure of the model seemed too risky and too artificial; keeping them external, as experimental assumptions whose implications could be compared to the 'free fall' cases and judged by their apparent explanatory power there, seemed more direct and honest; it keeps the suppositional nature of these issues in the foreground. Finally, when assumptions get embedded into a model, it becomes quite easy to forget that they are assumptions, and to embue them with an unwarranted reality (Hollnagel & Woods, 2006); by requiring them to be explicitly introduced by the experimenter, I hope to minimize that risk in the results which follow.

6.1 Fortuitous stopping

One way to begin approaching the issue of stopping the system as an adaptive strategy in overload would be to examine the behaviours that occur when "fortuitous stopping" occurs. Fortuitous stopping in this situation refers to a spontaneous and fortunate decrease in work demands. We have already seen the situation where, when near a tipping point, a small and brief increment in work demand forces a transition to a degraded state of operations from which the system is unable to recover (see page 7). The first experiment explores whether, and in what circumstances, can a fortuitous decrease in work demand following a pulse challenge that triggers such a

state change, be successful in reversing it. Understanding how a system might respond to a spontaneous decrement in demand should be useful in informing intentional stopping decisions.

Three aspects of "fortuitous stopping" are explored: the magnitude of the decrement (maximal decrease in the *New Work* rate); its duration (*ie*, do brief, high magnitude decrements differ appreciably in their effects from longer, low magnitude stops that are equivalent in their total demand); and the delay of the decrement (how closely it follows the initial pulse challenge that provoked the change in state).

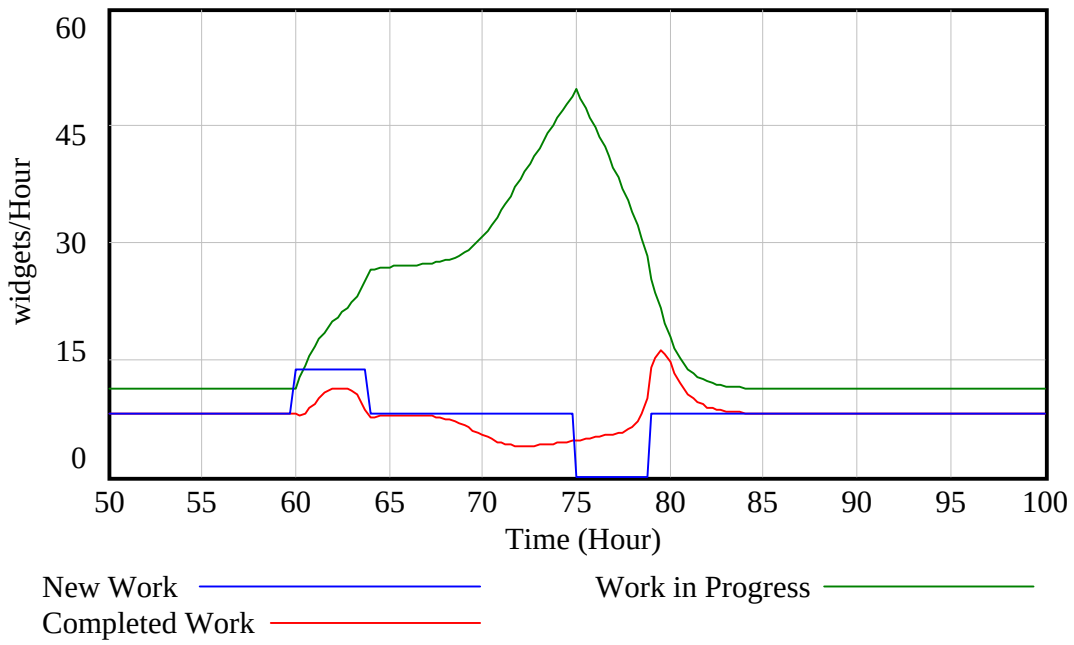


Figure 16. A sufficiently large and timely fortuitous decrement allows system recovery.

First, a sufficiently large and well-timed decrement in demand results in "flipping" the system back to its original state; the next two figures illustrate this behaviour. In Figure 16, the system is given the smallest pulse challenge of 4 hours duration that provokes a change in state from which it cannot spontaneously recover (in this particular scenario, that is an increase of ~71% over baseline). But, a fortuitous decrease in *New Work* of 4 hrs duration to 5% of baseline, occurring at hour 75 after the end of the pulse challenge is sufficient to allow the system to return to its original steady state. A smaller decrement, or one occurring later would produce only

temporary respite but would be insufficient to enable the system to recover and it would remain in its new, permanently degraded state of function (Figure 17, shows the effect of a decrement to 6% of baseline demand is inadequate to flip the system back to its original state).

If we examine these same two scenarios from the point of view of the compensatory resources that could be brought to bear in an overload crisis (represented by *Margin*) we can gain some insight into how the performance comes to be so different for such a small difference in the magnitude of the fortuitous decrease in demand.

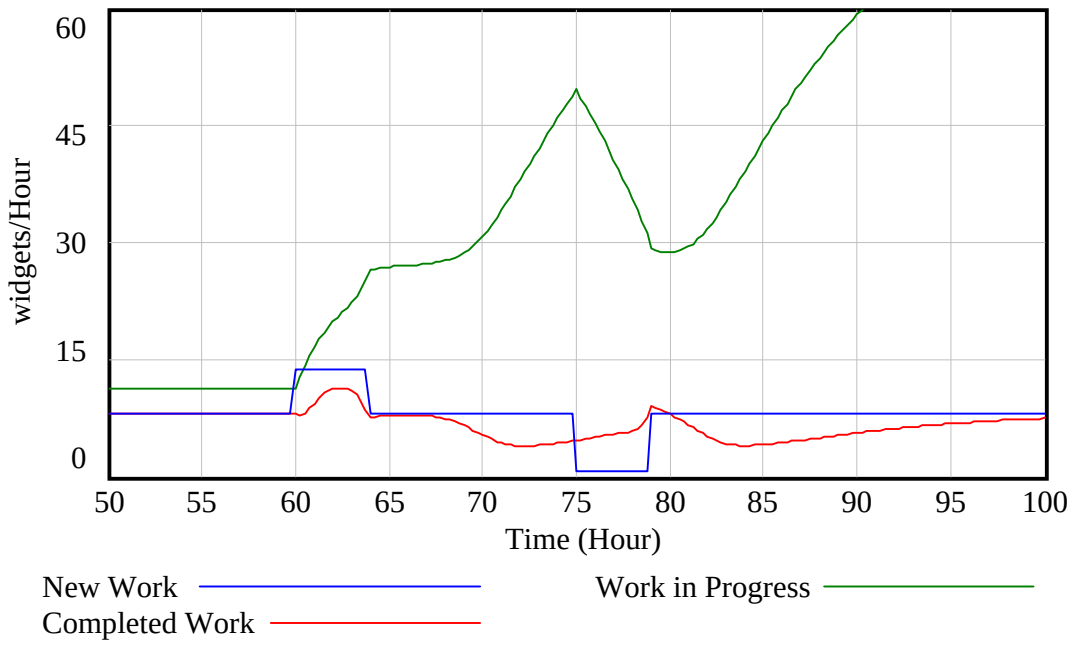


Figure 17. The ability to recover from a fortuitous decrement in demand critically depends on the magnitude of the decrease.

Figure 18 compares the behaviour of *Margin* in the same two scenarios of slightly differing fortuitous decrements in demand discussed above. The comparison shows that the difference in performance lies in the differences in the rates of restoration of these buffering resources. Consumption of *Margin* proceeds at identical rates in both situations before and immediately following the pulse challenge at hour 60. It falls to a low near zero at hour 76; this is analogous to an “all hands” situation in which all available resources are fully committed and there are no further reserves on which to

draw – a situation widely agreed to be avoided at all possible costs (Branlat & Woods, 2010). However, after the fortuitous decrements which start at hour 75, the two curves begin to diverge. Starting about hour 77, the restoration rate for the deeper decrement (to 5% of baseline) begins to increase at a slightly faster rate than does that for the smaller (to 6% of baseline) decrement, leading to a greater increase in *Margin* for the deeper decrement. Recall that the restoration causal loop is a positive feedback, *ie*, an amplifying, loop (see Figure 13). While positive feedback is often thought of as potentially destabilizing, here its ability to amplify small positive deviances is helpful (*cf* (Maruyama, 1963)); it triggers a virtuous rather than a vicious cycle that allows the system to claw its way back to normal.

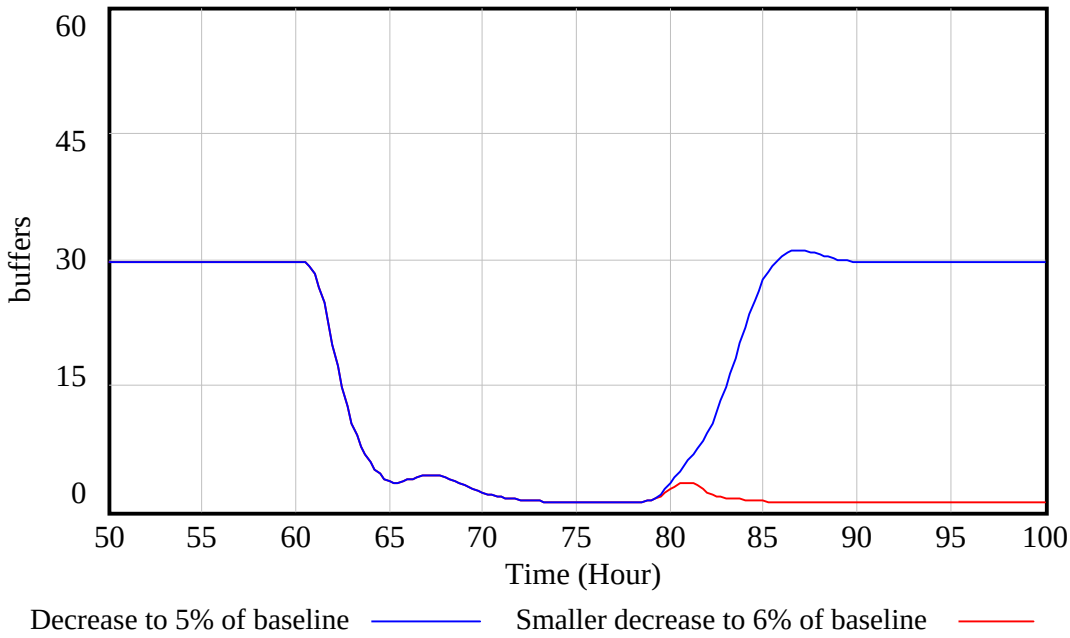


Figure 18. Effect of small changes in fortuitous decrease in demand on *Margin*.

There are potential implications for system design and resilience here. First, this finding suggests that one means to enhance a system's resilience (in the sense of recovery) would be to improve its ability to take advantage of, to capitalize on small fortuitous changes in the outside world (in this case, the incoming demand) when they occur. Second, it suggests two potential ways to do that. Once could develop ways to add directly to *Margin* when extremity demands it; it appears this could be effective

even if delivered sometime after the initial shock. Or, one might enhance the restoration loop, for example by decreasing the average restore time, or shortening the time to react to a critical depletion of *Margin.* (These possibilities are explored in Section 6.3).

Figure 19 illustrates the tradeoffs between the magnitude, duration, and timing of the smallest the decrement in demand that is able to trigger a return to steady state behaviour. It is apparent here that timing is a critical variable in recovery due to this sort of "input relief." If a decrement follows the pulse rather closely, it does not have to be very large or last very long; the longer the delay before this period of relief begins, the larger it must be and the longer it must last to be effective. Similarly, the magnitude of the decrease seems to be more important in triggering change than its duration. Note that the units of magnitude are widgets (work units) per hour, so the product of magnitude and duration is the total number of work units of which the system has been relieved, so for example a decrement of 87.5% for 6 hours is roughly equivalent (in total work units) to a decrement of 64% for 8 hours. Yet, the former will be effective up until 16 hours after the original challenge, but the later will have to occur at within at least 11 hours to be effective. An understanding of the tradeoffs among magnitude, duration, and timing of relief would seem to be informative in understanding how purposeful stopping might (or might not) be successful in recovery to normal operations.

In related experiments, I have explored the effects of the size, duration, and shape of the initial challenge. These explorations show that although pulse size does act in some ways as a non-linear, binary threshold (*ie*, below the threshold, the system will always recover spontaneously, but above the threshold value, it can never can), it still affects other properties. Below the threshold value, recovery time is affected by pulse magnitude as previously shown (see Figure 8). Above the threshold, the system always moves to a degraded state of functioning, but the final steady state value of *Work in Progress* is related to the magnitude of the pulse; however, pulse magnitude shows no relationship to the new steady state value of *Margin*, which descends close to zero and remains there once the threshold has been crossed (results not shown).

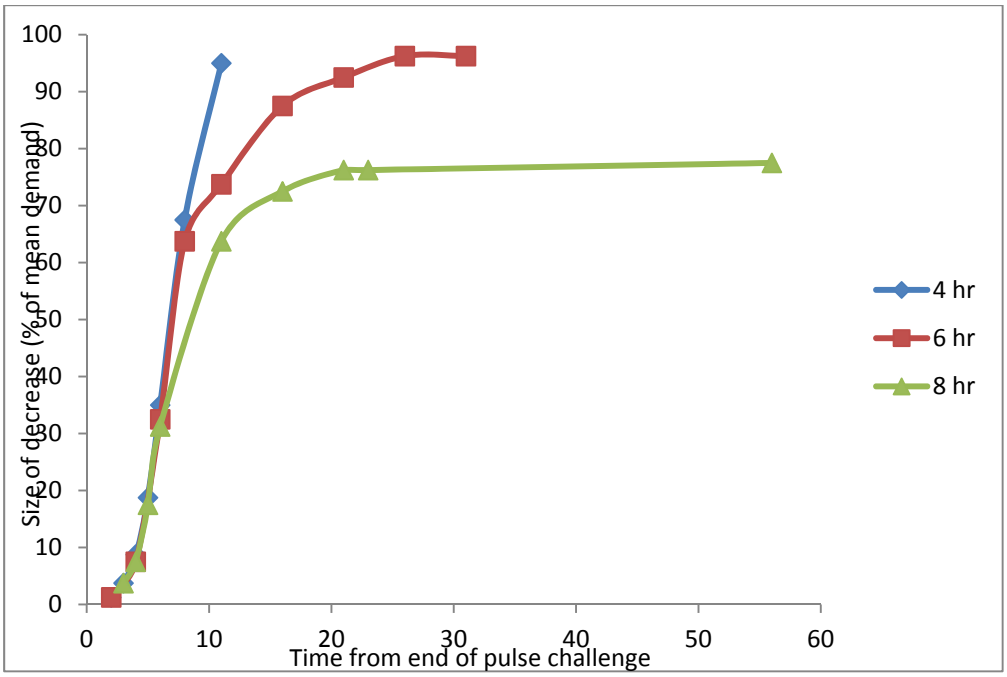


Figure 19. Relationship between magnitude, duration, and timing of the smallest decrement in demand able to trigger recovery.

There is an important difference between this sort of fortuitous relief and purposeful stopping in some systems (such as the ED), which highlights a fundamental difference between two classes of systems. In the fortuitous stopping scenarios just discussed, there is an assumption that the decrease in demand represents work that has simply vanished. While this might be the case for certain types of systems (*eg*, financial systems where stop trading procedures can ignore or reject incoming new orders), in other systems (such as the ED, flight or combat operations, or some process control systems), this work would instead be accumulated and have to be handled at some later time, thus potentially increasing the load on the system still further (Rudolph & Repenning, 2002). This difference will be explored later (see Section 6.3), but first it will be useful to examine a special case of fortuitous relief that is common to many work systems, including the ED: diurnal (or other regular) variations in workload.

6.2 Diurnal variation as fortuitous stopping

Demand in the ED and in many other systems follows well established regular patterns (daily, weekly, and seasonally, for example). The diurnal pattern of work in the ED (Welch, Jones, & Allen, 2007) is by far the strongest, and might function as a

regularly recurring instance of fortuitous stopping. This seems worth exploration since it could be the case that the recovery from an event such as 'free fall' might have been more due to the regular decrease in demand that occurs as evening wears on into night, rather than to some of the actions of the staff, in particular the stopping decision.

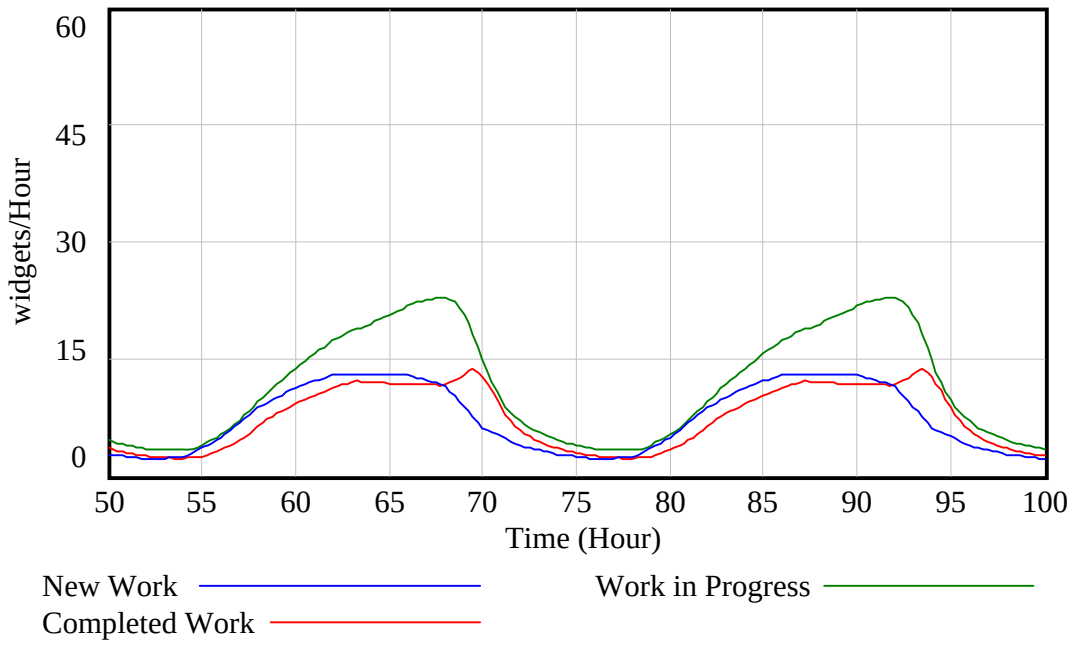


Figure 20. Diurnal variation in demand and workload.

Figure 20 shows the steady state cycle of system performance given a typical diurnal variation in *New Work*, where midnight occurs at 0, 24, 48, … hours, and noon as 12, 36, 60 hours, and so on. Note that workload (green line) and the rate of completed work (red line) tend to lag behind incoming work (blue line) in the expected manner. The diurnal pattern of incoming work used here is typical of ED systems (Welch, *et al.*, 2007); the peaks and valleys in this setting typically amount to roughly 70% over or under the mean rate, respectively, peaking around hours 64, 88, 112, … (corresponding to 1600 clock time), and reaching a low around hours 52, 76, 100, … (corresponding to 0400 clock time).

Now the timing of a pulse challenge becomes quite important. For example, a pulse arriving around noon (a fairly busy time, but before the typical workload peak), can

have an amplitude no greater than about 11% of the daily mean (in addition to the roughly 48% over mean experienced at noon) before triggering the same sort of shift to a degraded state of functioning.

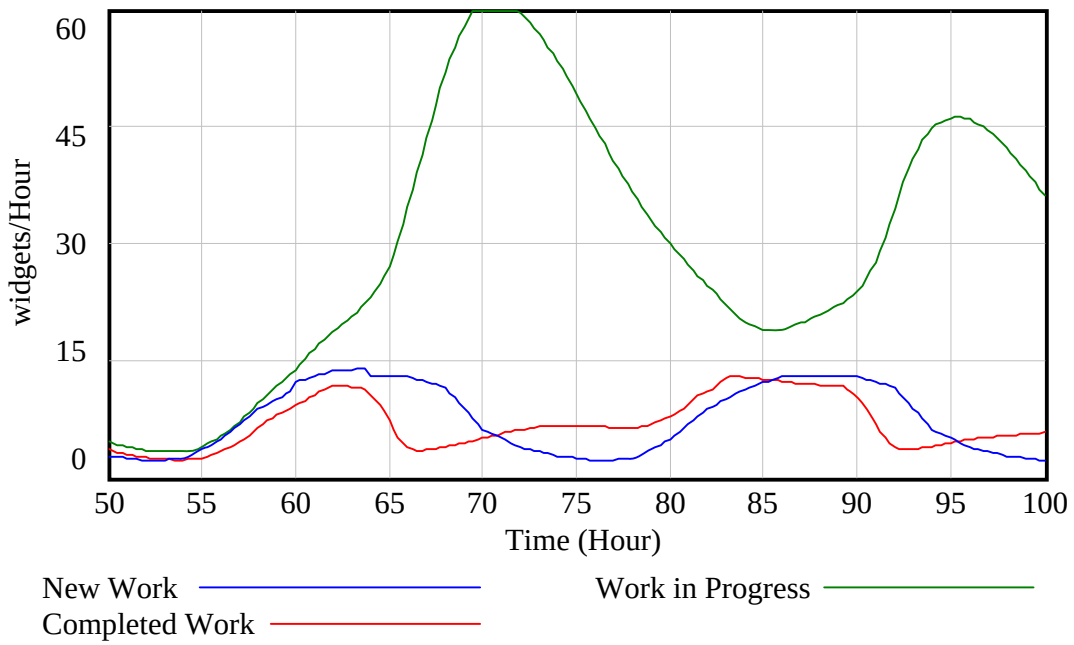


Figure 21. A small pulse challenge at hour 60 (noon) creates serious disruption but the system can ultimately recover.

Figure 21 shows behaviour at this point; there are four interesting points to note. First, the pulse challenge at hour 60 (corresponding to 1200 if translated to daily clock time) is barely discernable; in a real world system with a small amount of random variation, an impulse of this magnitude would be undetectable and its effects would seem to arise mysteriously, without obvious cause. This is exactly the experience reported in the 'free fall' cases. Second, disrupted operations carry over into the next day; the system does not return to its steady state cycle until almost hour 108, 48 hours after the beginning of the pulse challenge. Third, the peaks in workload (*Work in Progress*) now are significantly phase-shifted from the normal operations shown in Figure 20; under normal operations, *Work in Progress* peaks about hour 68 (corresponding to 2000 using 24 hour clock time), while after the pulse challenge, it peaks about hour 71 (corresponding to 2300, a time when workload is typically diminishing under normal conditions) for the next two days, before recovering its

normal pattern. Finally, the low in workload following the challenge much higher than normal and is only slightly less (~16%) than a normal peak (see Figure 22).

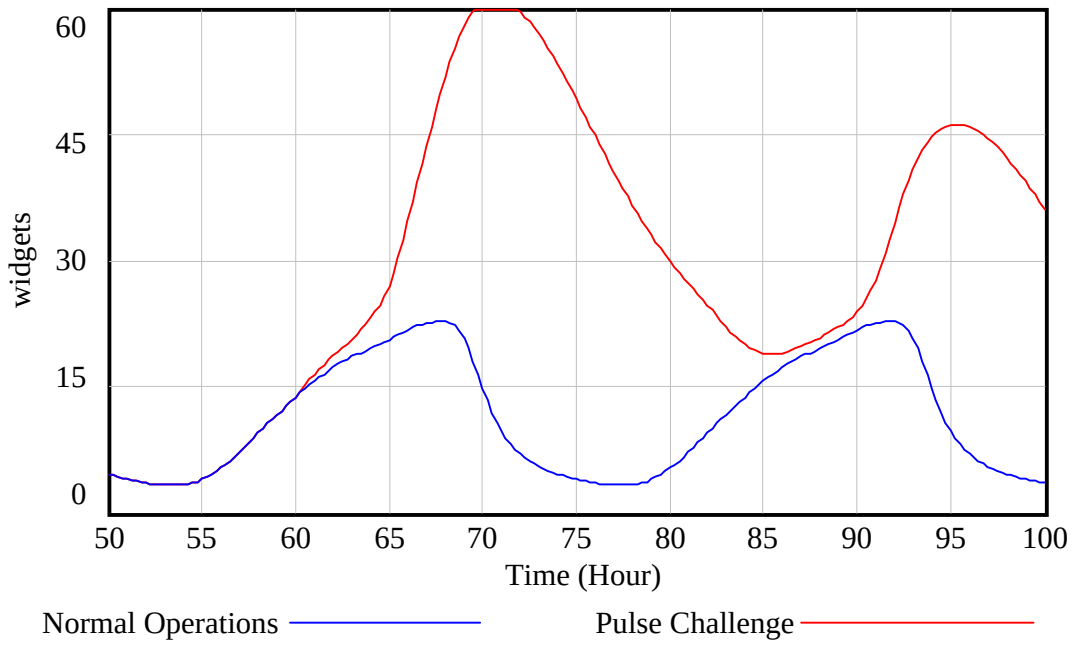


Figure 22. System load following a sub-threshold pulse challenge compared to normal operating conditions.

In comparison, a pulse challenge occurring at a regularly occurring low point in workload, such as hour 52 (corresponding to 0400 in the ED system) could be as large as 47% of mean without provoking a permanent shift to degraded operations (although with some disruption as noted above). Again, disruption persists for about 2 days, and is accompanied by a phase shift as previously noted.

Figure 24 shows the effect on *Margin* of a similar pulse demand, with a magnitude just beneath the system's "tipping point." (The time scale in these figures has been expanded to allow easier comparison to normal operations – the period preceding 60 hours; Figure 23 shows overall system performance in the same experiment for easy comparison). *Margin* collapses rapidly as the crisis develops, and does not return to its steady state value for 3 days. (It briefly approaches its normal value at about hour 107 (47 hours after the initial challenge) but it cannot be sustained).

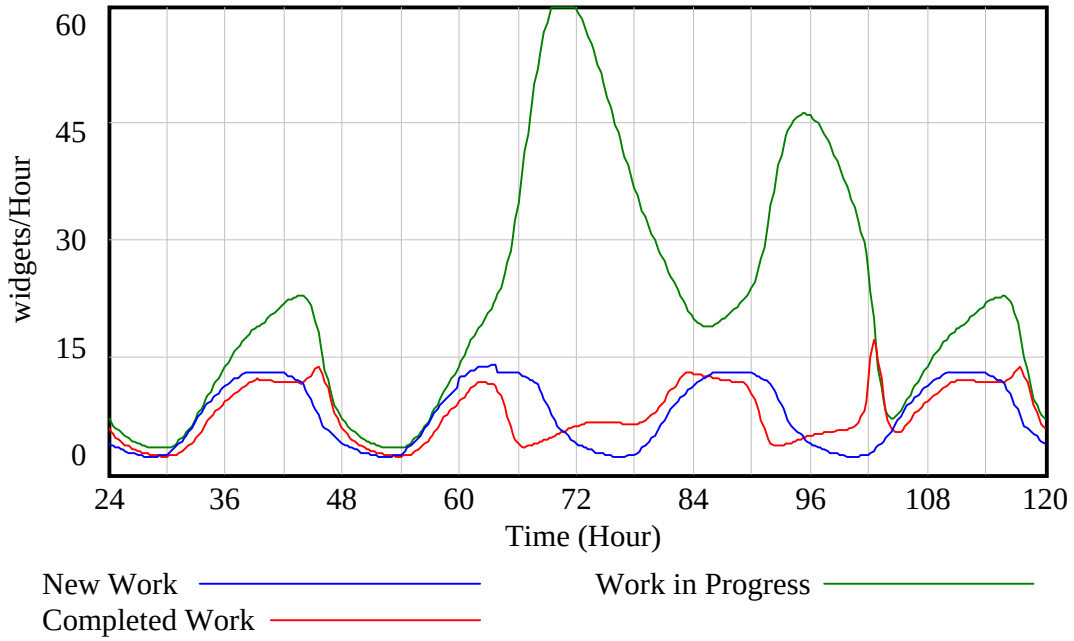


Figure 23. Response to subthreshold pulse challenge at hour 60.

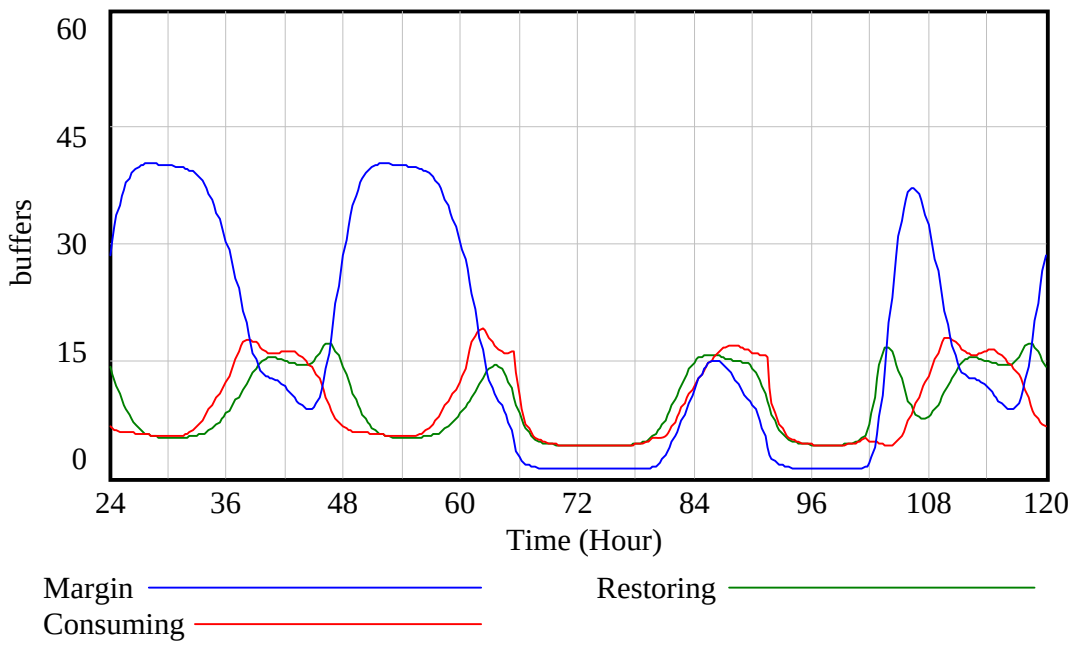


Figure 24. Effect on *Margin* of subthreshold pulse challenge at hour 60.

Further experiments showed qualitatively the same pattern noted for a non-pulsatile stream of demand: the potential for demand above some threshold value to flip the

system into a degraded operational state from which it could not recover; and tradeoffs among the magnitude, duration, and timing of fortuitous decreases in demand in terms of their effect of enabling the system to return to normal performance; and performance effects lagging behind and persisting after the initial challenging or rescuing pulses occur (data not shown).

Similarly, an analysis of *Margin* under the condition of diurnal variation of demand showed results similar to the steady state demand situation at the critical value of the fortuitous decrement; *Margin* was consumed at the same rate following the pulse challenge, but slightly different sizes of the fortuitous decrement in demand produced slightly different rates in *Restoring*, which was amplified by the positive feedback restoration loop to permit recovery if the decrement were sufficiently large.

This analysis of variation in demand as a form of fortuitous stopping raises important implications. There are potentially two opposing views of the role fortuitous stopping might play in enhancing or degrading resilient performance. The optimistic view would suggest that, by briefly "flirting" with the boundary of failure but being rescued by predictably occurring drops in demand, it might be possible for actors in a system to learn how better to deal with overload and what being close to the boundary feels like. But, to safely take advantage of this view one must be operating either in a fully controllable environment (such as a flight simulator, where subjects can practice in very close to "tipping point" conditions without risk to themselves or others), or the consequences of crossing the threshold and entering a downward spiral must be relatively benign and tolerable.

On the other hand, the pessimistic view suggests that the presence of fortuitous stopping in a system presents a risk, because it might easily breed overconfidence. Because system operators have experienced overload and been successful at regaining control due to fortuitous stopping, they may feel they are capable of handling additional demands by persevering, working a bit harder, or using a bit more margin. Thus, they do not learn to recognize when they are close to a threshold, and thus may cross the "tipping point" from which the system cannot recover, even after the excess demand (*eg*, the pulse challenge) has passed. Thus in systems that naturally experience regular variations in demand, it may be quite difficult to distinguish a safe

system from a much less safe system that has been “lucky” so far but in reality is teetering on the edge of collapse.

In summary, these two views suggest that in some carefully controllable situations, fortuitous stopping might contribute usefully to learning, but that in less controllable, more open systems, it is a potential hazard since actors in the system are likely to learn the wrong lessons from it.

In addition, traditional metrics of workload (such as peak load, *eg*, the highest values for *Work in Progress*) may not reliably indicate how close the system is coming to the point of collapse. However, the pattern of incomplete recovery and phase shift noted in response to just sub-threshold challenges might offer a potential signal of impending danger. In this situation, the system does not experience its normal recovery during the regularly expected periods of low demand; instead, the lowest values for *Work in Progress* are roughly commensurate with a normal high, and both the peaks and valleys of *Work in Progress* occur significantly later than expected. In particular, the valleys are more delayed than the peaks and occur at what are normally reasonably busy times (see Figure 21 and Figure 22).

This pattern resonates strongly with that reported in the “free fall” cases. In both cases, the overall level of demand (incoming patients, in this instance) was not appreciably greater than normal, but the morning shift (starting at 0700) began holding 25 patients, a load typical of mid-afternoon, not early morning. In such a circumstance, it may have taken only a small increase in load in the busy, late afternoon period (1500 and following) to cross the “tipping point” and enter the vicious cycle that was later described as “free fall.”

6.3 Purposeful stopping

I previously noted that fortuitous stopping has an advantage over purposeful stopping in many systems, since in purposeful stopping, incoming work not processed simply accumulates and adds to the total system load (Rudolph & Repenning, 2002). Hence, if purposeful stopping is to accomplish anything in an overload crisis, it must have some beneficial effect sufficient to counteract the malign effects of delay and increasing workload. In terms of the model presented here, that would take the form

of enabling an increase in *Margin*, either by adding additional resources, or by dramatically increasing the restoration rate. Such an increase in margin would generally come at some cost to the system, or else it would routinely be made available. For example, bringing in additional staff might be successful in meeting an overload demand, but would entail costs in overtime, fatigue, and burnout. In addition, it might actually spread the crisis through the larger organization, as it removes staff (or other resources) from other work they are already doing, thus posing a risk to the larger system in which this work system is embedded (*eg*, the hospital which contains the ED). This raises issues of level-crossing tradeoffs, local *vs* global optimization that are outside the scope of the current model. (Of course, this model could be extended to examine such situations: for example, coupling of systems, where the output of one is input to the next (*eg*, the ED and the ICU); or hierarchical situations, where several subsystems jointly contribute to the work of a super-system). For simplicity I chose not to represent those sorts of costs here, but assume that some increase in margin is available for extraordinary circumstances, but for a variety of reasons not represented in the model, these additional resources cannot be routinely or permanently, or perhaps even fully employed.

First we explore the effect of stopping as a means of restoring *Margin* – in effect, shifting priority from performing current work to replenishing the adaptive resources. This was done in a series of experiments in which the rate *Completed Work* was decreased, and at the same time, the rate *Restoring* was increased; the onset and duration of the stopping were separately controlled, as were the magnitudes of the changes in *Completed Work* and *Restoring* (*ie*, these could be separately varied). The results here seem to partially confirm but partially refute the impressions taken away from the "free fall" experience. First, the impression that stopping might be more effective if adopted sooner was supported. Figure 25 illustrates the effect of a one hour stop at hour 70 after a pulse challenge just over the "tipping point" in magnitude, lasting from hour 60 to 64. Here, the decrement in *Work Completed* was 50% of its then current value (*ie*, this was only partial stopping) and the smallest increase in *Restoring* sufficient to allow the system to return to near normal was substantial – almost 7 times greater than its then current value, to a level approximately 5 times its steady state value. Note also that the recovery here is incomplete. The system did not

return completely to its baseline steady state values, but stabilized in a new, modestly degraded condition, with *Work in Progress* about 28% higher, and *Margin* about 21% lower than before the pulse challenge. While this is not nearly as dramatic of a "phase shift" sort of degradation noted previously (where the increase in *Work in Progress* was roughly 400%; see Figure 8), it is still not a full return to normal. In addition, explorations of the possible values for the magnitudes of the restoration effort and work stoppage, and the rapidity with which they must be brought to bear to be effective, does not suggest this is likely to be a practically useful strategy.

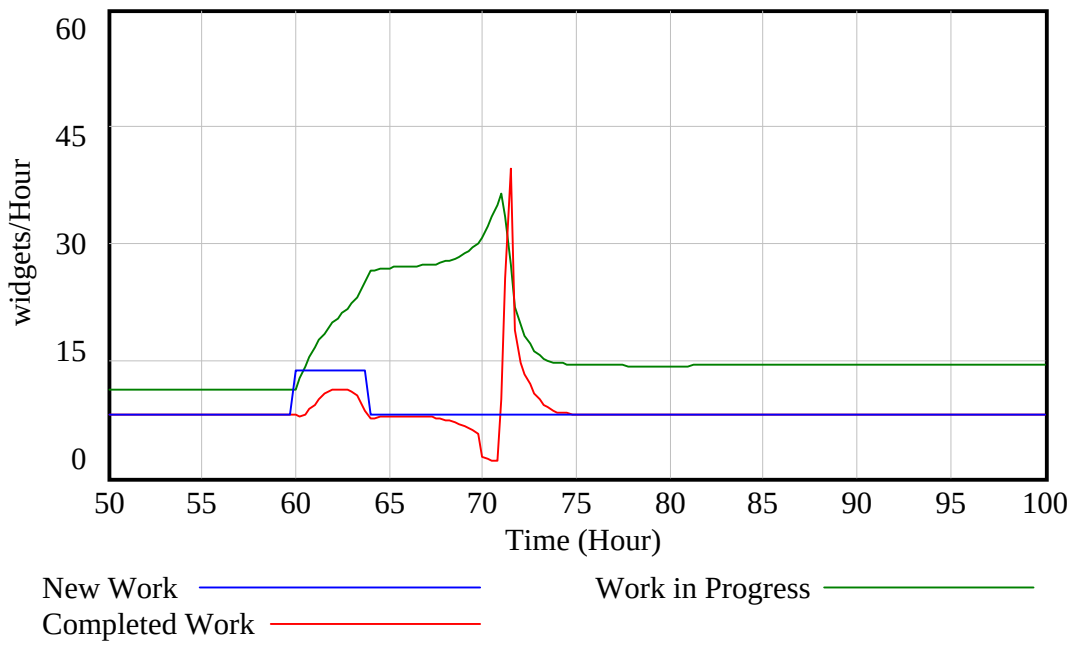


Figure 25. Effect of partial stopping to restore *Margin*.

Finally, I explored the effect of an exogenous infusion of additional margin resources in conjunction with stopping. Here again, although satisfactory combinations of variables could be found to successfully return the system to normal operations, they seemed impractical in that they required very rapid activation (*eg*, within 1 hour of the onset of the pulse challenge), and even minimal decreases in work output (*eg*, 10%) required disproportionately large additions of external resources. For example, with no stopping at all, an external addition to *Margin* of about 15% of baseline could effect a return to normal if applied quickly; but a 10% decrease in work output raised

the requirement to roughly 100%. This finding would imply that a small increase in the total of *Margin* resources available at all times might be more effective and possibly more efficient than having to mobilize larger increases on short notice.

French summary of Chapter 6

Ce chapitre décrit les expériences de simulation fonctionnant sur le modèle développé dans le chapitre 5. Les expériences impliquent des stratégies d'exploration, en particulier la stratégie d'arrêt (dans une certaine forme). Dans toutes les expériences décrites ici, un simple défi d'impulsion (une brève augmentation de nouveaux travaux, suivie d'un retour à la normale) a été utilisé.

La première série d'expériences examine l'arrêt fortuit - la situation où les apports (et donc la demande sur le système) tombe spontanément et permet donc potentiellement pour la récupération du système. Les expériences montrent que l'effet de l'arrêt fortuit dépend d'un compromis complexe entre trois facteurs: l'ampleur de la décroissance dans la demande; sa durée et son retard (c'est à dire, à quel point elle suit le défi impulsion initiale). Ces expériences ont montré que l'arrêt fortuite a un seuil de temps; plus le retard, plus le décrémenter doit être, jusqu'à un certain point, l'arrêt complet (ie, le taux d'entrée tombe à zéro) pour de longues durées est inefficace en permettant au système de remettre du choc tôt.

La première série d'expériences examine l'arrêt fortuit - la situation où les apports (et donc la demande sur le système) tombent spontanément et permettent donc potentiellement pour la récupération du système. Les expériences montrent que l'effet de l'arrêt fortuit dépend d'un compromis complexe entre trois facteurs: l'ampleur de la décroissance dans la demande; sa durée et son retard (c'est à dire, à quel point elle suit le défi impulsion initiale). Ces expériences ont montré que l'arrêt fortuit a un seuil de temps; plus le retard, plus le décrémenter doit être, jusqu'à un certain point, l'arrêt complet (i.e., le taux d'entrée tombe à zéro) pour de longues durées est inefficace en permettant au système de remettre du choc tôt.

La deuxième série d'expériences examinées naturellement des variations survenant dans la demande, telle que la variation diurne de la demande qui se produit dans les départements d'urgence. On peut penser le ralentissement de l'afflux qui survient pendant la nuit comme une forme de s'arrêter naturellement fortuite. Ces expériences ont montré que le moment de relever le défi pouls était critique; à des niveaux déjà élevés de flux, une petite impulsion presque indétectable pourrait déclencher un

changement d'un état dégradé à partir duquel le système n'a pas pu récupérer. (Un défi beaucoup plus grand survenant à des périodes normalement produisent une faible demande a été bien toléré). Ils ont également montré que, pour un défi à peine sous seuil (i e, un défi qui est juste en dessous du seuil qui déclencherait un changement d'état), perturbé les opérations ont continué pendant plusieurs cycles (par exemple, jours) après le défi était passée. En outre, les variations diurnes normales de la charge totale du système sont désormais déphasées, de sorte que la charge totale à des moments normalement bas était presque aussi élevée que la normale haute. Ce phénomène est remarquablement semblable à celle qui a été observée dans la "chute libre" des situations, et est donc une explication possible de leur origine. En outre, il pourrait bien servir comme un indicateur avancé que le système se rapproche d'un seuil critique.

La troisième série d'expériences examine l'arrêt délibéré pour permettre la restauration des ressources Marge consommée. Ils ont montré que l'arrêt délibéré est opportun, et peut réussir à déclencher une reprise des activités normales, si l'afflux ne s'accumule pas pendant la période d'arrêt (comme cela peut être le cas pour les systèmes financiers ou commerciaux). Toutefois, la reprise est souvent incomplète, le système s'améliore de son état dégradé, mais ne se reprend pas tout à fait du retour à la normale. Si des travaux neufs s'accumulent pendant l'arrêt, l'effet de l'arrêt délibéré est beaucoup plus contrasté. Bien que les combinaisons de variables pourraient être constaté que la récupération du système est activée, il semblait peu pratique en ce sens qu'ils nécessaire d'activation très rapide (par exemple, à 1 heure de l'apparition de relever le défi d'impulsion), et même des diminutions minimes dans la production de travail (par exemple, 10%) nécessaire ajouts disproportionnée des ressources externes pour être efficace.

Chapter 7. Inference and conclusion

This chapter discusses the implications of the preceding results in light of the original objective of the model. Although the model itself is much simpler than the simplest imaginable real world system, even this simple model demonstrates complex, non-intuitive and somewhat surprising behaviours. It seems unlikely that adding the missing complexity to this model to make it more realistic will make its performance any simpler or more intuitive.

This section begins by listing some of the more salient properties that the model and experiments with it have demonstrated, and moves to the more complex and potentially counterintuitive issues. The simpler and more "obvious" findings support the general utility of the model, and to the extent they match the observed variety of system performance, are a practical form of model validation (Hollnagel, 1993). Note that these general characteristics are likely to be present at multiple levels of any real world system.

7.1 Findings

The following items summarize the basic results gleaned from the foregoing experiments, roughly in order of simpler to more complex.

7.1.1 Nonlinearity

The behaviour of the system's main state variable (*Work in Progress*) cannot be captured by simple linear expressions such as "the more patients, the more crowded"; stimulus and response, input and output do not vary in proportional, or even commensurate, ways. And, neither is it captured by simple, smooth nonlinearities such as the "inverted U" curve, where performance first rises to the occasion, but then gradually declines. Rather, the collapses observed in these experiments occurred rather suddenly and without much warning, much as they did in the 'free fall' episodes.

7.1.2 Falling behind

Although the sorts of variables typically monitored in work systems (*ie*, input (*New Work*); workload (*Work in Progress*); and output (*Completed Work*)) do not provide very clear warnings of impending collapse, the time required to recover from a shock might provide such a warning and thus might well be worth monitoring. Particularly in systems with natural, regular cycles, such as the ED, the inability of periods of reduced demand (typically, night-time) to compensate for the overloads of the day – to "knit up the ravel'd sleave of care" (Shakespeare, ca 1606) as it were – might be reliable signals that the system is close to a threshold where even a small shock might tip it into a region of degraded performance from which it will be unable to recover. Thus, the model suggests the possibility of a quite practical key performance measure that could prove useful as a leading indicator of impending collapse.

7.1.3 Path dependence

The behaviour of the system cannot be entirely predicted from its current inputs and outputs, but is dependent on its recent history as well. Sufficiently large pulse challenges can produce changes in system performance that persist long after the challenge has past, and can only be reversed by some new exogenous or endogenous action. Another way to view this pattern, (and also the pattern of falling behind the tempo of events and delayed recovery) is to consider it as a form of hysteresis – a small input leads to a rapid and large change in behaviour, but when the input is reversed, the system does not retrace its previous path; rather, recovery lags behind for long (possibly infinite) time and requires a much larger input (in the opposite direction) to trigger it.

7.1.4 Threshold phenomena

All the experiments in this work (with the exception of the unrealistic, fully resilient system used for demonstration in Figure 5) demonstrate threshold behaviours, or "tipping points." These are values (typically for demand (*New Work*) but also seen for other variables) which when exceeded, produce a qualitative and permanent change in system performance; either the system spirals out of control in the case of complete collapse, or it shifts to a new steady state characterized by degraded

functioning (higher levels of *Work in Progress* and lower levels of *Margin*), and never recovers to its pre-shock level of performance.

Thresholds are common in complex adaptive systems, and have several implications relevant to this work and particularly to its motivating case studies. First, it seems likely that threshold points are inevitable in systems where the relative strength of positive and negative feedback loops determine performance. While it might be possible to modify the system so that its threshold is far from any reasonably anticipatable operating conditions, it is not possible to remove it altogether. Thus, the simple knowledge that thresholds must exist leads to the implication that to improve resilience, systems must have both the capability to recognize when they are nearing a threshold, and some means of breaking out of the "basin of attraction" (Alderson & Doyle, 2007, 2010; Carlson & Doyle, 2000; Csete & Doyle, 2002) into which they fall should they cross it.

Second, the existence of thresholds means that variability of demand or of resources can be an important issue because higher variance at same mean demand will eventually be more likely to produce a supra-threshold demand and trigger a state change. Thus, to the extent demand can be controlled, reducing the variability of demand might be a reasonable strategy for enhancing resilience. And, even if demand variability is not controllable, improving the ability to predict demand might at least enable systems to better cope with the unexpected. However, in open systems, an unexampled, extreme event is guaranteed to happen, given sufficient time (International Atomic Energy Agency, 2011; Snook, 2000); this suggests that a better strategy might be to assume variability will be higher than anticipated, that the distribution of demand will be heavy-tailed (Willinger, Alderson, Doyle, & Li, 2004), and to prepare accordingly. However, this raises classic, and rather difficult questions about system safety. How much resource should be devoted to protection *vs* production? How much margin is enough margin, especially when it is not often used? These questions are difficult because they entail no natural optimum, no point along the continuum which is unambiguously superior to any other point. But, knowing that the tradeoff exists, and understanding how variation (both controllable

and uncontrollable) affects risks might at least assist organizational leaders in managing their resource allocations.

7.2 Implications and application to 'free fall'

In this section, I connect the insights gained from this work back to the motivating examples, the 'free fall' situations. The 'free fall' episodes led to two fundamental questions (recall pg 7): how did it happen, and how was it recovered from?

7.2.1 A possible origin of 'free fall'

Why a 'free fall' episode ever should have occurred has been a puzzle. The number of patients arriving to the ED during these episodes was not particularly unusual; the experience of having several critically ill patients arrive in rapid succession was not at all unusual; the staff on duty in these episodes were well-trained and had long experience in the institution; and the issue of ED / hospital crowding was by no means a new one. So why did a crisis develop, when all the contributing factors had long been co-present in the organization without leading to catastrophic collapse?

The experiments that produced Figure 21 and Figure 22 provide some insight, by suggesting a set of conditions that might have led to a system collapse similar to 'free fall.' Figure 21 shows how a small, barely detectable random pulse, occurring at an already overloaded time, when the system is operating near a threshold, can create a disruption that persists for several days. This disruption moves the system closer to its threshold, and thus makes it more likely that a subsequent, even smaller disruption might cross the threshold and trigger a collapse.

This situation might be detectable in the phase shifting of workload; *eg*, an unusually high level of *Work in Progress* (analogous to the total number of patients in the ED) occurring at what would normally be a low point (in the ED, early in the morning, *eg*, 0400 – 0600). This is actually what occurred in the 'free fall' episodes, although its implications were not realized at the time. (The ED census in the ECC unit at 0700 on 14 December was 25, in a 22 bed treatment space; a typical 0700 census for this unit would be 10).

Thus, this analysis provides a possible explanation for the occurrence of the 'free fall' episodes. Crowding gradually moved the system closer to its threshold, making it more vulnerable to small random shocks; at the critical point, a small shock (something that might have been easily managed farther away from the threshold) kicked the system into a positive feedback vicious cycle, leading towards system collapse.

7.2.2 A possible means of recovery from 'free fall'

The second fundamental question raised by the 'free fall' episodes is what led to recovery? The general sense that developed from after-action review and reflection was that one means of regaining control was temporary stopping, and in particular, that stopping sooner might have decreased the period of danger to the system.

However, the experiments described in Sections 6.1 through 6.3 cast some doubt on this idea. Figure 19 shows that a small, beneficial decrement in demand, if fortuitously coordinated with the normally occurring decrement in demand associated with diurnal variation can be sufficient to push a degraded system back across the threshold in a favourable direction; but Figure 25 shows that purposeful stopping, when it is limited to stopping processing while incoming demand accumulates, is much less likely to produce this effect.

While it may be disappointing to confront the idea that the recovery from 'free fall' may have more to do with a fortuitous decrease in demand coupled with its normal diurnal slowing than it did to the extemporaneous invocation of a novel strategy (stopping to reorganize), there are still important implications for organizational design and management. Small decrements in demand may be common, but systems are not necessarily designed to be able to take advantage of them. Thus, enhancing the ability to capitalize on small fortuitous events might be an effective strategy. For example, reducing the lag between the development of a mismatch between capacity and demand and the mobilization of margin resources, or avoiding the "all hands" situation where all reserves are committed might be effective. Further exploration using the model might be useful in distinguishing the effectiveness of these strategies, or in particular, the circumstances favouring one or the other.

However beneficial enhancing the ability to capitalize on favourable changes in demand might be, it is important to realize that this strategy is most effective in a pulse challenge – a temporary increase in demand that eventually resolves on its own. Exploratory experiments (data not shown) on a step or ramp challenge (a slow or rapid, permanent increase in demand) suggest a strategy of opportunistic capitalization would not be as effective, and other methods of dealing with the challenge must be sought.

Another idea that arose in after-action reflection following the 'free fall' episodes was that monitoring some aspect of *Margin* resources – either the absolute level, the rate of consumption or rate of restoration – might provide an early warning that the system was nearing a critical point. In particular, the thinking was that since restoration typically seems optional and deferrable, decreases in restoration activity (for example, stocking, in an ED) might be used to signal impending problems. But, the experiment resulting in Figure 24 shows that although a decrease in *Restoration* does accompany an overload crisis, it tends to parallel *Work in Progress*, and thus is not likely to be useful as an early indicator of problems.

These findings suggest that the wrong lessons may have been learned from the 'free fall' episodes. Repeated experiences with approaching the critical threshold but then recovering (*eg*, due to diurnal variation) might falsely lead to an understanding that specific strategies, not the regular, simple, night-time decompression, were sufficient to manage episodes of overload. Thus, stopping may have been ineffective in comparison to regular decompression but misinterpreted as an effective strategy[19].

7.2.3 *A possible strategy to reduce the risk of 'free fall'*

Taken together, these results suggest a rough form of a possible strategy to enhance resilience in an overload crisis such as resulted in 'free fall'. By *monitoring* periods of naturally occurring recovery (*eg*, the early AM ED census), it might be possible to

---

[19] Of course, stopping may have had other beneficial effects in recovery beyond simple, mechanistic effects on workload. It may still have been useful in taking advantage of a fortuitous respite. It may have encouraged staff to press on when feeling overwhelmed, similar to the effect of the erroneous map in the semi-apocryphal story of a military unit lost in the Alps using a map to reach safety, only to discover then that their map was of the Pyrenees (Weick, 2001).

anticipate periods of vulnerability, and *adapt* by preemptively increasing *Margin* before a crisis is apparent, rather than having to react quickly, when it is already full-blown and risk being too late to be effective. Second, additional adaptations in the form of shortening the delay between recognition of mismatch and deployment of margin resources to allow the system to take opportunistic advantage of fortuitous decreases in demand is also likely to be effective. Both strategies have the advantage of not requiring a permanent, sustained increase in *Margin* (although that too might be effective, albeit expensive).

Interestingly, an attempt to put this strategy into practice is just now beginning in the organization that suffered the 'free fall' episodes. The hospital had begun a regular "bed meeting" on days when crowding or bed shortages were experienced; typically, these were held around 1600, because at that time, the number of new admissions and planned discharges were relatively well known, and it gave sufficient lead time to adjust the nighttime nursing staff levels for the following evening (nurses work 0700 – 1900, or 1900 – 0700 shifts in most of the hospital). This timing may have worked reasonably well for the hospital as a whole, but by 1600 the fate of the ED was (in most cases) already determined, and actions taken at the bed meeting were often "too little, too late" to have much effect in decompressing the ED. Because of this, the hospital has now agreed that, on days when the ED is already holding an unusually large number of patients at 0800, a quick bed meeting will be held at 0900 to see if a small number of beds for ED patients might be freed up. If this action can actually be taken, then it would provide an empirical test of the strategy of providing smaller amounts of resources early when nearing a threshold, and of the utility of the phase shift and failure of night-time recovery as an indicator of impending crisis.

Over the longer term, the recognition that workers in the ED system may tend to learn the wrong lessons from their occasional brushes with overload should serve as a cautionary tale and would help foster the "preoccupation with failure" noted in the high reliability organisation literature (Rochlin, 1999; Rochlin, *et al.*, 1987; Weick, Sutcliffe, & Obstfeld, 1999). It will be admittedly difficult to keep this focus.

While these results cannot be viewed as conclusive, they are certainly sufficient to force a re-examination of what could or should be learned from these experiences.

7.3 What was gained by modeling

These results demonstrate the value of the modeling exercise. The value is not the specific results, but rather that the process of modeling and analyzing the results leads to new thinking about the problem. That is, a literal, direct application of the model results is not warranted, given the simplifications and assumptions involved; rather, the model afforded a means of structured thinking about the problem and the conditions that contributed to it, and thus led to new insights and revision of previous thinking about 'free fall'.

In addition to explaining 'free fall', and to re-thinking the effectiveness of the adaptations used in it, the modeling exercise suggests some possibilities for enhancing the system's resilience. Although Figure 25 suggests that a "just in time" strategy of adding additional *Margin* resources in an overload crisis is not likely to be effective, there are other forms of "just in time" strategies that, instead of emphasizing "front end" analysis and preparation – anticipate everything that will happen or that will be needed in advance – instead focus on developing general problem-solving skills, expanding the repertoire of skills, and supporting quick decision-making (Weick, 2001).

7.4 Limitations

"All models are wrong; some models are useful" (Box, 1976).

Building a theory, or a model, is a seductive and potentially dangerous exercise, for four reasons (Cilliers, 2001). First, models, by their very nature, always make some kinds of sacrifices, and it is difficult to know in advance whether or not those sacrifices are consequential (Dekker, 2011). Any reasonably useful model tends to becomes 'second nature', and its simplifications and assumptions quickly forgotten (Hollnagel & Woods, 2006). Second, modelers frequently become enamoured of their creations, and thus reluctant to subject them to the sort of challenge that might undermine the entire enterprise. Third, when the system is characterized by non-linearities (as this one is), then it is fundamentally impossible to assess the importance of elements which have been left out. Fourth, when the system exhibits path

dependence (as is the case here), if the model's history and the modeled system's history are not kept identical, then the model and the system will diverge.

These problems are exacerbated by the genesis of the motivating example in a healthcare setting, because, in contrast to several other theory- (model-) generating explorations of critical incidents (Snook, 2000; Vaughan, 1996), the volume of empirical data typically available in healthcare settings after accidents or critical events is extremely scanty. I attempted to compensate for this relative paucity of data by moving across settings, examining case studies of other events, in other settings, and looking for isomorphisms among them (J G March, Sproull, & Tamuz, 1991). And, I have attempted to maintain an approach of fitting the model to the data, rather than fitting the data to the model (Le Coze, 2008).

The remainder of this section attempts to list the principle assumptions and simplifications that are embedded in this model, so that its use will at least be informed and the risk of "mistaking the map for the territory" can be minimized.

7.4.1 Simplifications

Compared to the systems that inspired it, this model is extremely simplified in two specific ways. First, it represents only a single level of behaviour and of control, but most real world systems have multiple levels that mutually interact, and also often operate on different time scales. To make things worse, these hierarchies need not have well-developed, clearly nested structures, but levels may interpenetrate each other such that it can be difficult to characterize subordinate and superordinate positions (Cilliers, 2001). While the behaviours seen here might be reasonably expected to be manifest in any of these other levels, the problems of cross-scale interaction are entirely bounded out of the current exercise. Further, it would not be reasonable to take an extremely high level view (*ie*, that this model subsumes the behaviours at lower levels and represents on the aggregate, the resultant of these effects) because such a view implicitly adopts the reductionist stance that high level properties are obtainable by simply combining all the lower level ones. But properties that are emergent (such as resilience or safety) cannot be derived from their components; thus any reference from these results to a complex, real world system

would have to address the question of whether failing to include cross level effects was important for the question at hand and carefully consider its disadvantages (Hollnagel & Woods, 2006).

Second, even at a single level in a hierarchy, the output of many work systems is often the input to another work system at the same level; that is, they form production chains rather than simple input – throughput – output systems. (For example, the output of the ED in terms of patients is the input to the hospital wards, the intensive care units, the operating room, and the outpatient care system, and so on). There are many examples of complex interactions among the components of these chains; in fact the old term "ED overcrowding" is being replaced by "ED / hospital overcrowding" in healthcare management circles in a direct reflection of this knowledge. Extraordinarily complex behaviours have been noted in supply chain models, which are closely but imperfectly related to hospital production chains (Mosekilde & Laugesen, 2007).

Several additional simplifications are readily apparent. Demand is aggregated into a simple, one-dimensional variable (*New Work*), but of course demand in real world systems is a multi-dimensional vector, a collection of different sorts of variables with different values (*eg*, patients with different conditions, degrees of severity and urgency). Resources used to maintain margin of maneuver are similarly aggregated into a simple one-dimensional variable (*Margin*), while in reality these resources might be quite diverse (*eg*, space, workers, procedures, short-cuts, buffers, strategies) and more important, some sorts of resources might be well suited to certain types of demand but ineffective in others. Finally, the output of the system is similarly simplified to a one-dimensional variable (*Completed Work*), while real out is multidimensional (*eg*, patients, completeness and quality of the work performed, timeliness, economy of effort, risk incurred or avoided, *etc*). This simplification is compounded by the simple relationship of overload to performance, since overload in real systems may differentially affect different dimensions of system performance through a complex set of relationships including strategies and actions engaged in by actors in the system.

7.4.2 System dynamics assumptions

Finally, the choice of system dynamics as a modeling formalism brings with it important assumptions and limitations that may not be immediately apparent. System dynamics methods focus on aggregate performance measures, the "central tendency" as it were of system variables under a variety of conditions. This yields important insights, but in many circumstances, the properties of interest are not in the center of the distribution of performance but in the tails – *eg*, an ED that performs well on the vast majority of patients but extremely poorly for one or two cannot be thought of as performing well in the same sense as, say, a jelly-bean factory with a similar performance pattern.

Second, system dynamics methods carry an implicit assumption of continuity and infinite divisibility, in that as a stock decreases, it might become vanishingly small but never reaches zero. But, in real world systems resources such as *Margin* are often discrete – as an ED consumes its temporary bed locations, the number of additional spaces remaining decreases in discrete jumps, not in smooth fractional decreases, and ultimately reaches zero. These "structural zeros" may provoke additional discontinuities and state changes in real systems that would not be captured in a model such as this one. Similar effects relating to treating discrete entities as continuous can be noted with non-zero boundaries as well. However, by aggregating many separate entities (some continuous and some discrete) into a composite variable such as *Margin*, this effect should be minimized in the current model[20].

Finally, in view of the forgoing limitations, any claims made for the results should be similarly limited. Thus, this thesis does not claim that the dynamics noted here will occur, but rather only that they do occur, and uses the model to explore and document the conditions that contribute to them (Axelrod & Cohen, 2000).

---

[20] In fairness, there are methods to account for this effect in the system dynamics armamentarium, but they add additional complexity and tend to bind the model more tightly to the specific system being modeled.

7.5 Future Directions

Although the work to this point has proven useful in stimulating reflective thought about how systems respond resiliently or brittlely to an overload challenge, there is still a large potential for future explorations.

A natural extension of this work would be to make it less abstract and refine the model so that it reflects ED work much more specifically. The resultant model is much more detailed, but shows the same general dynamics reported above: threshold phenomena; delayed recovery from a small shock when the system is operating near its threshold; and strong time dependence in the effect of adding additional resource in response to a crisis (*ie*, resources added quite early can abort a system collapse, but the same or larger infusion of resources added late are ineffective). A paper resulting from this extension is included in Appendix 3; the figures in this paper illustrate those behaviours. This work also supports a degree of face validity of the model, in that the behaviours seen in this extension appropriately express those seen in the ED.

A very important "next step" will be to apply and/or extend the model to at least one other domain. Since the modeling exercise originated in an ED crisis, even though the intent was to develop a general method for thinking about resilience in many, not-yet-specified complex work systems, its origins in the ED may account for a substantial portion of its apparent applicability to that domain. Such an attempt would of course require partnership with researchers with deep domain expertise in the new area of application, but would be extremely important in establishing the generalizability of this approach.

Finally, the experiments discussed above could be rather easily extended to examine a large number of possibilities flowing rather logically from the previous explorations, which might have implications for operation in actual practice.

For example, it might be useful to modify the consumption of *Margin* to explicitly avoid (or at least reduce the possibility of) an "all hands" situation (one in which all available resources are fully committed). This might be particularly interesting, as many real world systems try to follow this practice (Branlat & Woods, 2010); slowing the consumption of these resources may not adversely affect workload too severely

(since they are already quite depleted), but might allow a system to recover more quickly, or to be better positioned to take advantage of a smaller random, fortuitous decrements in demand. Thus there is a potential tradeoff between current production and future resilience (recovery) that could be explored further (Hoffman & Woods, 2011; Woods & Branlat, 2011b).

Second, a common response to extreme overload situations is to shift effort from restoration to production; but if that allowed resources to drop to a critically lower value than it would have otherwise, such a strategy might reduce the ability of the system to recover in the future. This would be another form of the current production *vs* future recovery tradeoff noted above. Interestingly, it partially contradicts the thinking in the previous paragraph – when resources are already low, is it important or immaterial that they drop no farther? A modeling approach is a useful way to sort out such questions, not so much in the sense that one notion is right and the other wrong (although that could be the case), but rather to identify the circumstances and situations which contribute to making one or the other idea more suitable as a guide.

Third, preliminary explorations not presented in detail here have suggested that the rapidity of response to overload (both in consumption of margin resources and in their restoration) might be critically important in both mitigating the disruption caused by overload and in recovering to normal operations afterward, so a systematic exploration of this behaviour might be useful.

Finally, the current model might be usefully extended in ways that would facilitate exploration of the tradeoff between exploitation and exploration more explicitly. A bias towards exploitation might entail lower short run risk but reduced adaptability, while a bias towards explorations might produce either a high risk / high reward circumstance, or even a "churning" effect that would be deleterious. Thus future work might entail an examination of the balance between exploration and exploitation, and in particularly whether dynamically shifting that balance (*eg*, as a form of "breakout" strategy when the system become trapped in a degraded state) might be advantageous.

French summary of Chapter 7

Ce dernier chapitre résume les principales conclusions et les implications de la thèse, il discute des limitations, et termine avec quelques réflexions sur les orientations des travaux futurs.

Les expériences montrent que les performances de ce système peuvent être caractérisées par la:

- Non-linéarité. Le comportement des variables du système principal (Work in Progress) ne peut pas capturée par de simples expressions linéaires telles que «plus les patients, le plus encombré». Et, il n'est ni capturé par de simples non-linéarités lisses telles que le «U inversé» courbe, où la performance s'élève d'abord à l'occasion, mais ensuite diminue progressivement. Plutôt, les effondrements observés dans ces expériences se sont produits plutôt soudainement et sans avertissement beaucoup, beaucoup comme cela s'est produit en «chute libre».
- Le retard le rythme des opérations. L'incapacité des périodes de demande réduite (typiquement, la nuit) pour compenser la surcharge de la journée pourrait être des signaux fiables que le système est proche d'un seuil où même un petit choc peut faire basculer dans une région de la dégradation des performances à partir de laquelle il sera incapable de récupérer. Ainsi, le modèle laisse entrevoir la possibilité d'une mesure de performance clé très pratique qui pourrait se révéler utile comme indicateur avancé de l'effondrement imminent.
- Chemin de la dépendance. Le comportement du système peut ne pas être entièrement prédit à partir de ses entrées et sorties actuelles, mais dépend de son histoire récente aussi. Défis d'impulsion suffisamment grande peut produire des changements dans la performance du système qui persistent longtemps après le défi a passé, et ne peut être renversée par une nouvelle action exogène ou endogène.
- Phénomènes de seuil. Toutes les expériences dans ce travail montrent des comportements de seuil, ou «points de basculement.» Ce sont des valeurs pour

les variables critiques, qui, lorsqu'on les dépasse, produisent un changement qualitatif et permanent des performances du système; soit le système de spirales hors de contrôle dans le cas d'effondrement complet, ou qu'il se déplace vers un nouvel état stationnaire caractérisé par un fonctionnement dégradé et ne se remettra jamais de son niveau de performance pré-choc.

Ces résultats ont des implications importantes pour la compréhension «chute libre». L'origine de ces épisodes a été un casse-tête. Le nombre de patients arrivant aux urgences au cours de ces épisodes n'a pas été particulièrement inhabituel, l'expérience d'avoir plusieurs patients gravement malades arrivent en succession rapide n'était pas du tout inhabituelle, le personnel de service ont été bien formés et ont une longue expérience dans l'établissement; et la question du SAU / hôpital surpeuplement n'était pas une nouvelle. Alors pourquoi est-ce qu'une crise se développe, quand tous les facteurs qui contribuent depuis longtemps dans l'organisation, sans effondrement catastrophique? Les expériences réalisées ici suggèrent un ensemble de conditions qui auraient pu conduire à un effondrement du système similaire à «chute libre». Pendant les périodes surchargées, petites, à peine perceptibles provoquent des chocs et perturbent les opérations qui durent pendant plusieurs cycles (par exemple, 2 jours ou plus dans un SAU) après le choc est passé. Pendant ces périodes, le système est particulièrement vulnérable dans cette désorganisation encore plus petite pourrait le pousser à travers le seuil et déclencher un effondrement du système. Rétrospectivement, cette attitude peut être vue dans les précédents de la "chute libre" épisodes. Ainsi, cette analyse fournit une explication possible de la survenue d'épisodes de la «chute libre». Le surpeuplement progressivement déplacé le système plus proche de son seuil, ce qui rend plus vulnérables aux petits chocs aléatoires; au point critique, un petit choc (quelque chose qui aurait pu être facilement géré plus loin du seuil) a lancé le système dans un cycle vicieux de réactions positives , conduisant à l'effondrement du système.

Deuxièmement, ces résultats remettent en question le sens général qui s'est développé à partir après action et de réflexion que l'on veut dire de reprendre le contrôle a été l'arrêt temporaire, et en particulier, que l'arrêt plus tôt pourrait avoir diminué la période de danger pour le système, depuis l'arrêt délibéré pour le moins un niveau

raisonnable ou la durée ne serait pas suffisant pour aboutir à la récupération, tout en arrêtant fortuite pourrait facilement le faire. Ainsi arrêt peut avoir été associés à la récupération dans l'esprit des acteurs dans le système pour de mauvaises raisons; arrêt n'a pas facilité la récupération, mais la récupération (via la diminution de la demande diurne normale) ne facilitent stopping.ing vers l'effondrement du système.

Pris ensemble, ces résultats suggèrent une forme grossière d'une stratégie possible pour améliorer la résilience en période de crise telle que la surcharge a entraîné «chute libre». En périodes de surveillance de la récupération naturelle (par exemple, au début des années AM SAU recensement), il pourrait être possible d'anticiper les périodes de vulnérabilité et à s'adapter en augmentant préventivement marge avant qu'une crise se manifeste, plutôt que d'avoir à réagir rapidement, quand il est déjà à part entière et risque d'être trop tard pour être efficace. Deuxièmement, des adaptations supplémentaires sous la forme d'un raccourcissement du délai entre la reconnaissance de l'inadéquation et le déploiement des ressources de marge pour permettre au système de profiter des opportunistes diminue de fortuit dans la demande est également susceptible d'être efficace. Les deux stratégies ont l'avantage de ne pas nécessiter une permanente, augmentation soutenue de la marge (bien que trop peut être efficace, mais cher). Fait intéressant, une tentative pour mettre cette stratégie en pratique commence tout juste à l'organisation qui a subi des épisodes de la «chute libre».

Il existe plusieurs limites à ce travail. Premièrement, les modèles, par leur nature même, toujours faire quelques sortes de sacrifices, et il est difficile de savoir à l'avance si oui ou non ces sacrifices sont corrélatives (Dekker, 2011). Tout modèle est raisonnablement utile tend à «seconde nature» devient, et ses simplifications et hypothèses vite oubliés (Hollnagel & Woods, 2006). Deuxièmement, la genèse de ce travail dans un établissement de santé peut avoir une incidence favorable sur le match dans les comportements entre le modèle et l'exemple de motivation. Troisièmement, le modèle implique quelques simplifications drastiques: flux, les sorties de la demande de travail, et les ressources sont tous représentées aussi simple, unidimensionnel quantités continues, quand dans tout système réel, ils sont des agrégations hétérogènes. Quatrièmement, le système réel est souvent composé de

groupes de sous-systèmes, où la sortie de l'un est l'entrée à l'autre, ou l'un est subsumé par un autre dans une hiérarchie. Ce modèle ne peut pas résoudre les problèmes qui pourraient résulter de l'accouplement ou des passages à niveau présents dans de tels cas, même si elle pourrait éventuellement être étendue à explorer ces situations. Toutes ces questions (et plus) signifient que, toute réclamation faite par les résultats devrait être aussi limitées. Ainsi, cette thèse ne prétend pas que la dynamique noter ici va se produire, mais seulement qu'ils ne se produisent, et utilise le modèle à explorer et à documenter les conditions qui contribuent à leur (Axelrod et Cohen, 2000).

Enfin, les travaux sur la dynamique de résilience n'est que maintenant que commencer. Les travaux futurs pourraient utilement envisager quelques-uns des domaines suivants. Une extension naturelle de ce travail serait de la rendre moins abstraite et affiner le modèle afin qu'il reflète le travail SAU beaucoup plus spécifiquement. Le modèle qui en résulte est beaucoup plus détaillé, mais montre la même dynamique générale rapportée ci-dessus: phénomènes de seuil; reprise retardée d'un petit choc lorsque le système est près de son seuil, et la dépendance du temps fort dans l'effet de l'ajout de ressources supplémentaires en réponse à une la crise (par exemple, les ressources ajoutées très tôt peut interrompre un effondrement du système, mais la même perfusion ou plus des ressources ajoutées tardivement sont inefficaces). Un document résultant de cette extension est inclus dans Appendice 3.

Une prochaine étape très importante sera d'appliquer et / ou étendre le modèle à au moins un autre domaine. Depuis l'exercice de modélisation origine dans une crise SAU, même si l'intention était de développer une méthode générale pour la réflexion sur la résilience dans de nombreux, non encore précisé systèmes de travail complexes, son application à certaines nouvelles zones serait extrêmement importante dans l'établissement de la généralisabilité de cette approche.

References

Adams, J. G., & Biros, M. H. (2001). The endangered safety net: establishing a measure of control. *Acad Emerg Med, 8*(11), 1013-1015.

Alderson, D. L., & Doyle, J. C. (2007, 7-10 Oct. 2007). *Can complexity science support the engineering of critical network infrastructures?* Paper presented at the Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on.

Alderson, D. L., & Doyle, J. C. (2010). Contrasting Views of Complexity and Their Implications For Network-Centric Infrastructures. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 40*(4), 839-852.

American College of Emergency Physicians. (1990a). Hospital and emergency department overcrowding. *Ann Emerg Med, 19*(3), 336.

American College of Emergency Physicians. (1990b). Measures to deal with emergency department overcrowding. *Ann Emerg Med, 19*(8), 944-945.

Anders, S., Woods, D. D., Wears, R. L., Perry, S. J., & Patterson, E. S. (2006). *Limits on adaptation: modeling resilience and brittleness in hospital emergency departments*. Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France,8 - 10 November 2006.

Argyris, C., & Schön, D. (1974). *Theory in Practice: Increasing Professional Effectiveness*. London, UK: Jossey-Bass.

Ashby, W. R. (1957). *An Introduction to Cybernetics*. London, UK: Chapman & Hall Ltd.

Axelrod, R., & Cohen, M. D. (2000). *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. New York, NY: Basic Books.

Baer, R. B., Pasternack, J. S., & Zwemer, F. L., Jr. (2001). Recently discharged inpatients as a source of emergency department overcrowding. *Acad Emerg Med, 8*(11), 1091-1094.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791 - 799.

Branlat, M., Anders, S., Woods, D. D., & Patterson, E. S. (2008). Detecting an erroneous plan: does a system allow for effective cross-checking? In E. Hollnagel, C. P. Nemeth & S. W. A. Dekker (Eds.), *Resilience Engineering Perspectives: Remaining Sensitive to the Possibility of Failure* (Vol. 1, pp. 247 - 257). Aldershot, UK: Ashgate.

Branlat, M., & Woods, D. D. (2010). *How do Systems Manage Their Adaptive Capacity to Successfully Handle Disruptions? A Resilience Engineering*

Perspective. Paper presented at the AAAI Fall Symposium 2010, Arlington, VA,11 - 13 November 2010. http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2238

Buck, S. J. (1992). Book review: Governing the Commons. *Natural Resources Journal, 32*, 415 - 417.

Carlson, J. M., & Doyle, J. (2000). Highly Optimized Tolerance: Robustness and Design in Complex Systems. *Physical Review Letters, 84*(11), 2529.

Carlson, J. M., & Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences, 99*(Supplement 1), 2538-2545. doi: 10.1073/pnas.012582499

Cilliers, P. (2001). Boundaries, Hierarchies and Networks in Complex Systems. [Article]. *International Journal of Innovation Management, 5*(2), 135.

Congressional Budget Office. (2001). Causes and Lessons of the California Electricity Crisis (pp. 52). Washington, DC: US Congress.

Cook, R. I. (2006). *A very simple resilience definition? (No!) Model? (No!) Example!* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Cook, R. I., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. *Quality & Safety in Health Care, 14*(2), 130-134.

Crossan, M. M., & Apaydin, M. (2010). A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature *Journal of Management Studies* (Vol. 47, pp. 1154-1191).

Csete, M. E., & Doyle, J. C. (2002). Reverse Engineering of Biological Complexity. *Science, 295*(5560), 1664-1669.

Cuvelier, L., & Falzon, P. (2011). Coping with uncertainty: resilient decisions in anaesthesia. In E. Hollnagel, J. Pariès, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice: A Guidebook* (pp. 29 - 43). Farnham, UK: Ashgate.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation models. *Academy of Management Review, 32*(2), 480 - 499.

Dekker, S. W. A. (2005). Why we need new accident models (pp. 11). Ljungbyhed, Sweden: Lund University School of Aviation.

Dekker, S. W. A. (2011). *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems*. Farnham, UK: Ashgate.

Dekker, S. W. A., & Woods, D. D. (1999). To Intervene or not to Intervene: The Dilemma of Management by Exception. *Cognition, Technology & Work, 1*(2), 86-96. doi: 10.1007/s101110050035

Derlet, R., Richards, J., & Kravitz, R. (2001). Frequent overcrowding in U.S. emergency departments. *Acad Emerg Med, 8*(2), 151-155.

Derlet, R. W., & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Annals of Emergency Medicine, 35*(1), 63-68.

Dijkstra, E. W. (2008). Forum: (A look back at)GoTo statement considered harmful. *Communications of the ACM, 51*(1), 7 - 8.

Flin, R. (1996). *Sitting in the Hot Seat: Leaders and Teams for Critical Incident Management*. New York, NY: Wiley.

Forrester, J. W. (1985). “The” model versus a modeling “process”. *System Dynamics Review, 1*(1), 133-134. doi: 10.1002/sdr.4260010112

Fu, W.-T. (2007). Adaptive tradeoffs between exploration and exploitation: a rational-ecological approach. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 165 - 179). Oxford, UK: Oxford University Press.

Garfinkel, H. (1967). "Good" organizational reasons for "bad" clinic records *Studies in Ethnomethodology* (pp. 186 - 207). Cambridge, UK: Blackwell Publishing Ltd.

Gilardi, S., Guglielmetti, C., Perry, S. J., Pravettoni, G., & Wears, R. L. (2009). *Changing horses in mid-stream: sudden changes in plan in dynamic decision problems*. Paper presented at the 9th International Naturalistic Decision-Making Conference, Covent Garden, London, UK,4 - 6 June 2007.

Goldberg, C. (2000, December 17, 2000). Emergency crews worry as hospitals say, 'No vacancy', *New York Times*, pp. Section 1, pg 27.

Goldfarb, Z. A. (2010, 21 May 2010). SEC launches inquiry into market's "flash crash". *Washington Post*, from http://www.washingtonpost.com/wp-dyn/content/article/2010/05/20/AR2010052005086.html

Gordon, J. A., Billings, J., Asplin, B. R., & Rhodes, K. V. (2001). Safety net research in emergency medicine: proceedings of the Academic Emergency Medicine Consensus Conference on "The Unraveling Safety Net". *Acad Emerg Med, 8*(11), 1024-1029.

Heylighen, F., Cilliers, P., & Gershenson, C. (2007). Complexity and Philosophy. In J. Bogg & R. Geyer (Eds.), *Complexity, Science and Society* (pp. 117 - 134). Oxford, UK: Radcliffe Publishing.

Hodges, J. S. (1991). Six (Or So) Things You Can Do with a Bad Model. *Operations Research, 39*(3), 355-365.

Hoffman, R. R., & Woods, D. D. (2011). *Simon's slice: five fundamental tradeoffs that bound the performance of all human work systems*. Paper presented at the 10th International Conference on Naturalistic Decision Making, Orlando, FL,31 May - 3 June 2011.

Holling, C. S. (1973). Resilience and Stability of Ecological Systems. *Annual Review of Ecology and Systematics, 4*(1), 1-23. doi: doi:10.1146/annurev.es.04.110173.000245

Hollnagel, E. (1993). Requirements for dynamic modelling of man-machine interaction. *Nuclear Engineering and Design, 144*(2), 375-384. doi: 10.1016/0029-5493(93)90153-z

Hollnagel, E. (2004). *Barriers and Accident Prevention*. Aldershot, UK: Ashgate.

Hollnagel, E. (2009). *The ETTO Principle: Efficiency-Thoroughness Tradeoff (Why Things That Go Right Sometimes Go Wrong)*. Farnham, UK: Ashgate.

Hollnagel, E. (2011). Prologue: the scope of resilience engineering. In E. Hollnagel, J. Pariès, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice: A Guidebook* (pp. xxix - xxxiv). Farnham, UK: Ashgate.

Hollnagel, E., Cacciabue, P. C., & Hoc, J.-M. (1995). Work with technology: some fundamental issues. In J.-M. Hoc, P. C. Cacciabue & E. Hollnagel (Eds.), *Expertise and Technology: Cognition and Human-Computer Cooperation* (pp. 1 -18). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hollnagel, E., & Sundström, G. (2006). States of resilience. In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 339 - 346). Aldershot, UK: Ashgate.

Hollnagel, E., & Woods, D. D. (2006). Epilogue: resilience engineering precepts. In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 347 - 358). Aldershot, UK: Ashgate.

Hunte, S. G. (2010). *Creating Safety in an Emergency Department.* PhD, University of British Columbia, Vancouver. Retrieved from https://circle.ubc.ca/handle/2429/27485

Institute of Medicine. (2006). Hospital-Based Emergency Care At the Breaking Point. In T. N. A. Press (Ed.). Washington, D.C.: Institution of Medicine of the National Academies.

International Atomic Energy Agency. (2011, 2 June 2011). Fukushima Nuclear Accident Update Log Retrieved 24 June 2011, from http://www.iaea.org/newscenter/news/tsunamiupdate01.html

. Job Switching. (1952). *I Love Lucy Episode Guide* Retrieved 24 July 2011, from http://cgi.pathfinder.com/time/time100/artists/video/lucyfactory.mov

Johnson, S. (2010). *Where Good Ideas Come From: The Natural History of Innovation*. New York, NY: Riverhead Books.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice - Hall.

Kauffman, S. A. (1995). *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford, UK: Oxford University Press.

Kauffman, S. A. (2000). *Investigations*. Oxford, UK: Oxford University Press.

Kelen, G. D., Scheulen, J. J., & Hill, P. M. (2001). Effect of an emergency department (ED) managed acute care unit on ED overcrowding and emergency medical services diversion. *Acad Emerg Med, 8*(11), 1095-1100.

Kellermann, A. L. (2000). Déjà vu. *Ann Emerg Med, 35*(1), 83-85. doi: S0196064400188158 [pii]

Kellermann, A. L. (2006). Crisis in the Emergency Department. *N Engl J Med, 355*(13), 1300-1303. doi: 10.1056/NEJMp068194

Kellermann, A. L. (2010). Waiting room medicine: has it really come to this? *Ann Emerg Med, 56*, in press.

La Porte, T. R. (1996). High reliability organizations: unlikely, demanding, and at risk. *Journal of Contingencies and Crisis Management, 4*(2), 60 - 71.

Le Coze, J.-C. (2008). Disasters and organisations: From lessons learnt to theorising. *Safety Science, 46*(1), 132-149.

Lengnick-Hall, C. A., & Beck, T. E. (2009). Resilience capacity and strategic agility: prerequisites for thriving in a dynamic environment. In C. P. Nemeth, E. Hollnagel & S. W. A. Dekker (Eds.), *Resilience Engineering Perspectives: Preparation and Restoration* (Vol. 2, pp. 39 - 69). Aldershot, UK: Ashgate.

Little, J. D. C. (1961). A proof of the queuing formula: L = AW. *Operations Research, 9*(3), 383 - 387.

Lundberg, J., & Johannson, B. (2006). *Resilience, stability and requisite interpretation in accident investigation*. Paper presented at the 2nd Symposium on Resilience Engineering, Juan-les-Pins, France,8 - 10 November 2006. http://www.resilience-engineering.org/REPapers/Lundberg_Johansson.pdf

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71 - 87.

March, J. G., Sproull, L. S., & Tamuz, M. (1991). Learning from samples of one or fewer. *Organization Science, 2*(1), 1 - 13.

Maruyama, M. (1963). The second cybernetics: deviation-amplifying mutual causal processes. *American Scientist, 5*(2), 164 - 179.

May, A. D. (1989). *Traffic Flow Fundamentals*. Englewood Cliffs, NJ: Prentice Hall.

Miller, J. G. (1960). Information input overload and psychopathology. *American Journal of Psychiatry, 116*(8), 695 - 704.

Mosekilde, E., & Laugesen, J. L. (2007). Nonlinear dynamic phenomena in the beer model. *System Dynamics Review, 23*(2-3), 229-252.

Nemeth, C. P. (2009). The ability to adapt. In C. P. Nemeth, E. Hollnagel & S. W. A. Dekker (Eds.), *Resilience Engineering Perspectives: Preparation and Restoration* (Vol. 2, pp. 2 - 12). Aldershot, UK: Ashgate.

Ostrom, E., Eggertsson, T., & Calvert, R. (Eds.). (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. New York, NY: Cambridge University Press.

Pariès, J. (2011). Resilience and the ability to respond. In E. Hollnagel, J. Pariès, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice: A Guidebook* (pp. 3 - 8). Farnham, UK: Ashgate.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies.* New York, NY: Basic Books.

Perrow, C. (1994). The Limits of Safety: The Enhancements of a Theory of Accidents. *Journal of Contingencies & Crisis Management, 2*(4), 212.

Piaget, J. (1967). *Biologie et Connaissance*. Paris, France: Gallimard.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science, 27*(2/3), 183 - 213.

Reeder, T. J., & Garrison, H. G. (2001). When the safety net is unsafe: real-time assessment of the overcrowded emergency department. *Acad Emerg Med, 8*(11), 1070-1074.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155 - 169.

Roberts, K. (1989). Research in nearly failure-free, high reliability organizations: having the bubble. *IEEE Trans Engineering Management, 36*, 132-139.

Rochlin, G. I. (1999). Safe operation as a social construct. *Ergonomics, 42*(11), 1549 - 1560.

Rochlin, G. I., La Porte, T. R., & Roberts, K. H. (1987). Self-designing high reliability: aircraft carrier flight operations at sea. *Naval War College Review, 40*(4), 76-90.

Roe, E., & Schulman, P. R. (2008). *High Reliability Management: Operating on the Edge*. Stanford, CA: Stanford Business Books.

Rudolph, J. W., Morrison, J. B., & Carroll, J. S. (2009). The dynamics of action-oriented problem solving: linking interpretation and choice. *Academy of Management Review, 34*(4), 733 - 756.

Rudolph, J. W., & Repenning, N. P. (2002). Disaster Dynamics: Understanding the Role of Quantity in Organizational Collapse. *Administrative Science Quarterly, 47*(1), 1-30.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., *et al.* (2009). Early-warning signals for critical transitions. *Nature, 461*(7260), 53-59.

Schneider, S., Zwemer, F., Doniger, A., Dick, R., Czapranski, T., & Davis, E. (2001). Rochester, New York: a decade of emergency department overcrowding. *Acad Emerg Med, 8*(11), 1044-1050.

Schull, M. J., Szalai, J. P., Schwartz, B., & Redelmeier, D. A. (2001). Emergency Department Overcrowding Following Systematic Hospital Restructuring: Trends at Twenty Hospitals over Ten Years. *Acad Emerg Med, 8*(11), 1037-1043.

Schulman, P. R. (1993a). Analysis of high reliability organizations: a comparative framework. In K. H. Roberts (Ed.), *New Challenges to Understanding Organizations*. New York, NY: Macmillan.

Schulman, P. R. (1993b). The negotiated order of organizational reliability. *Administration and Society, 25*(3), 353 - 372.

Schulman, P. R., & Roe, E. (2007). Designing Infrastructures: Dilemmas of Design and the Reliability of Critical Infrastructures. *Journal of Contingencies and Crisis Management, 15*(1), 42-49.

Schulman, P. R., Roe, E., Eeten, M. v., & Bruijne, M. d. (2004). High Reliability and the Management of Critical Infrastructures. *Journal of Contingencies and Crisis Management, 12*(1), 14-28. doi: 10.1111/j.0966-0879.2004.01201003.x

Shakespeare, W. (ca 1606). The Tragedie of Macbeth (pp. Act 2, Scene 2, line 45). London, UK.

Shanker, T., & Richtel, M. (2011, 17 January 2011). In the new military, data overload can be deadly, *New York Times*. Retrieved from

https://www.nytimes.com/2011/01/17/technology/17brain.html?pagewanted=all

Snook, S. A. (2000). *Friendly Fire: The Accidental Shoot-down of US Black Hawks over Northern Iraq*. Princeton, NJ: Princeton University Press.

Stephens, R. J. (2010). *Managing the margin: a cognitive systems engineering analysis of emergency department patient boarding.* PhD, The Ohio State University, Columbus, OH.

Stephens, R. J., Woods, D. D., Branlat, M., & Wears, R. L. (2011). *Colliding dilemmas: interactions of locally adaptive strategies in a hospital setting*. Paper presented at the 4th International Conference on Resilience Engineering, Sophia Antipolis, France,6 - 8 June 2011.

Sterman, J. D. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin McGraw-Hill.

Sutcliffe, K. M., & Weick, K. E. (2008). Information overload revisited. In W. H. Starbuck & G. P. Hodgkinson (Eds.), *Oxford Handbook of Organisational Decision Making* (pp. 56 - 75). Oxford, UK: Oxford University Press.

Sweeney, J. L. (2002). *The California Electricity Crisis*. Stanford, CA: Hoover Institution Press.

US General Accounting Office. (2003). Hospital Emergency Departments: Crowded Conditions Vary Among Hospitals and Communities (pp. 71). Washington, DC: US General Accounting Office.

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*. Chicago, IL: University of Chicago Press.

von Bertalanffy, L. (1973). *General System Theory* (revised ed.). New York, NY: George Braziller.

Voß, A., Procter, R., Slack, R., Hartswood, M., & Rouncefield, M. (2006). Understanding and supporting dependability as ordinary action. In K. Clarke, G. Hardstone, M. Rouncefield & I. Sommerville (Eds.), *Trust in Technology: A Socio-Technical Perspective* (pp. 195 - 216). Dordrecht, NL: Springer.

Wears, R. L. (2010). Health information technology risks. *The Risks Digest, 26*(25). Retrieved from http://catless.ncl.ac.uk/Risks/26.25.html#subj1

Wears, R. L., & Perry, S. J. (2007). Status boards in accident & emergency departments: support for shared cognition. *Theoretical Issues in Ergonomics Science, 8*(5), 371 - 380. doi: http://dx.doi.org/10.1080/14639220701194304

Wears, R. L., Perry, S. J., Anders, S., & Woods, D. D. (2008). Resilience in the Emergency Department. In E. Hollnagel, C. P. Nemeth & S. W. A. Dekker

(Eds.), *Resilience Engineering: Remaining Sensitive to the Possibility of Failure* (pp. 193 - 210). Aldershot, UK: Ashgate.

Wears, R. L., Perry, S. J., & McFauls, A. (2006). *Free fall - a case study of resilience, its degradation, and recovery, in an emergency department*. Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France. http://www.resilience-engineering.org/REPapers/Wears_et_al.pdf

Wears, R. L., Perry, S. J., & McFauls, A. (2007). *Dynamic changes in reliability and resilience in the emergency department*. Paper presented at the 51st Human Factors and Ergonomics Society, Baltimore, MD,October 2007.

Wears, R. L., Perry, S. J., & Nasca, L. (2007). *'Free fall' - highly decentralized, resilient adaptation to demand-capacity mismatches in an emergency department*. Paper presented at the 8th International Naturalistic Decision-Making Conference, Pacific Grove, CA,4 - 6 June 2007.

Wears, R. L., Perry, S. J., Salas, E., & Burke, C. S. (2005). Status boards in emergency departments: support for shared cognition. In R. Tartaglia, S. Bagnara, T. Bellandi & S. Albolino (Eds.), *Healthcare Systems, Ergonomics and Patient Safety* (pp. 273 - 280). Leiden, NE: Taylor & Francis.

Wears, R. L., & Webb, L. K. (2011). *Fundamental on situational surprise: a case study with implications for resilience*. Paper presented at the 4th International Conference on Resilience Engineering, Sophia Antipolis, France,6 - 8 June 2011.

Weick, K. E. (1987). Organizational culture as a source of high reliability. *California Management Review, 29*(2 (Winter 1987)), 112-127.

Weick, K. E. (1990). The vulnerable system: an analysis of the Tenerife air disaster. *Journal of Management, 16*(3), 571-596.

Weick, K. E. (1993). The collapse of sense-making in organizations: the Mann Gulch disaster. *Administrative Science Quarterly, 38*(4 (Dec 1993)), 628 - 652.

Weick, K. E. (2001). *Making Sense of the Organization*. Malden, MA: Blackwello.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (1998). *Organizing for high reliability: processes of collective mindfulness*. Paper presented at the Proceedings of the Second Annenberg Conference on Enhancing Patient Safety and Reducing Errors in Health Care, Rancho Mirage, CA.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (1999). Organizing for high reliability: processes of collective mindfulness. In B. Staw & R. Sutton (Eds.), *Research in Organizational Behavior: Volume 21* (Vol. 21, pp. 81 - 123). Greenwich, CT: JAI Press.

Welch, S. J., Jones, S. S., & Allen, T. (2007). Mapping the 24-Hour Emergency Department Cycle to Improve Patient Flow. *Joint Commission Journal on Quality and Patient Safety, 33*, 247-255.

Willinger, W., Alderson, D., Doyle, J. C., & Li, L. (2004). *More "normal" than normal: scaling distributions and complex systems*. Paper presented at the 36th Winter Simulation Conference, Washington, D.C.

Woods, D. D. (2011). Resilience and the ability to anticipate. In E. Hollnagel, J. Pariès, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice: A Guidebook* (pp. 121 - 125). Farnham, UK: Ashgate.

Woods, D. D., & Branlat, M. (2010). Hollnagel's test: being 'in control' of highly interdependent multi-layered networked systems. *Cognition, Technology & Work, 12*(2), 95-101. doi: 10.1007/s10111-010-0144-5

Woods, D. D., & Branlat, M. (2011a). Basic patterns in how adaptive systems fail. In E. Hollnagel, J. Paries, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice* (pp. 127 - 144). Farnham, UK: Ashgate.

Woods, D. D., & Branlat, M. (2011b). *How human adaptive systems balance fundamental trade-offs: Implications for polycentric governance architectures*. Paper presented at the 4th International Conference on Resilience Engineering, Sophia Antipolis, France,6 - 8 June 2011.

Woods, D. D., Dekker, S. W. A., Cook, R. I., Johannesen, L., & Sarter, N. (2010). *Behind Human Error* (2nd ed.). Farnham, UK: Ashgate.

Woods, D. D., & Sarter, N. B. (2000). Learning from automation surprises and "going sour" accidents. In N. B. Sarter & R. Amalberti (Eds.), *Cognitive Engineering in the Aviation Domain* (pp. 327 - 353). Hillsdale, NJ: Lawrence Erlbaum Associates.

Woods, D. D., & Shattuck, L. G. (2000). Distant Supervision–Local Action Given the Potential for Surprise. *Cognition, Technology & Work, 2*(4), 242-245.

Woods, D. D., & Wreathall, J. (2008). Stress-Strain Plots as a Basis for Assessing System Resilience. In E. Hollnagel, C. P. Nemeth & S. W. A. Dekker (Eds.), *Resilience Engineering: Remaining Sensitive to the Possibility of Failure* (pp. 143 - 158). Aldershot, UK: Ashgate.

Woods, D. D., Wreathall, J., & Anders, S. (2006). *Stress-strain plots as a model of an organization's resilience.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Zhou, T., Carlson, J. M., & Doyle, J. (2005). Evolutionary dynamics and highly optimized tolerance. *Journal of Theoretical Biology, 236*(4), 438-447.

Zijlstra, F. R. H., Roe, R. A., Leonova, A. B., & Krediet, I. (1999). Temporal factors in mental work: effects of interrupted activities. *Journal of Occupational and Organizational Psychology, 72*(2), 163 - 185.

Zink, B. J. (2006). *Anyone, Anything, Anytime: A History of Emergency Medicine*. Amsterdam, NL: Elsevier.

Zwemer, F. L. (2000). Emergency Department Overcrowding. *Annals of Emergency Medicine, 36*(3), 279.

Appendix 1 Methods

This section summarizes the methods used to investigate and analyse the 'free fall' incidents, and to develop empirical data to inform the modelling exercise.

The two 'free fall' incidents were investigated using the critical incident method of debriefing participants. Since I was a participant in one of the episodes, this was a combination of participant and non-participant observation. The critical incident method basically involved 3 passes through the incident: first to establish a general time line and identify points of potential interest or intervention; second to examine those points with respect to what subjects knew, what they noticed, what their goals and understandings were; and third to revisit those points in the form of alternatives, what might they have done differently, how might a novice (or expert) failed (or succeeded) in these conditions. These sessions were conducted as one on one interviews with nurses, physicians, and techs involved in the incidents. The staff involved directly in the two incidents did not overlap; typically there were 2 – 3 physicians and 2 – 3 nurses as subjects for each incident. These data were supplemented by examination of ED logs (containing patient arrival and dispositions times), and by attendance at after-action reviews conducted by the organization in response. Data were recorded as field notes; the interviews were not audio-recorded.

Further information about the performance of the ED under overload conditions was gathered by a non-clinical systems analyst (JBM), who was not a member of the care provider organization, as part of an NIH funded project to study the effects of overcrowding in acute coronary syndrome ("heart attack") patients on which I served as principal investigator (National Heart Lung and Blood Institute grant number 1 R21 HL098875-01). These data were captured in field notes.

A grounded theory approach using the constant comparative method was used to elicit thematic understandings of the nature of crowding and the ED's response to it.

Appendix 2 Model equations

The model was implemented in Vensim Professional v. 5.10e (Ventana Systems). The equations used in the model are documented here. They are grouped into stocks, flows, and auxiliary variables.

The stocks, or accumulations:

```
Work in Progress= INTEG (
New Work Rate-Completed Work Rate, New Work Rate*av time to
completion)
Units:      widgets
The variable represents total work in the system -- "in progress"
here is used to encompass both work actually being done, and work
which is queued up, waiting for some resource to be freed.  This
model does not explicitly represent either of those processes.

Margin= INTEG (
Restore Rate-Consume Rate, 33)
Units:      buffers
This variable represents additional resources that can be brought to
bear to deal with overload.  The specific nature of those resources
is not specified: they may be buffers (hence the units), but need not
be, as they could also be additional staff, procedural resources (eg,
authorization to skip some steps), unofficial resources (tacitly
supported task shedding, etc).
```

The flows into or out of the stocks:

```
New Work Rate=
8*24*table for diurnal variation(MODULO(Time,24))
Units:      widgets / Hour
This variable represents demand -- rate at which new work is added to
the system.
For the steady state (non-varying) arrivals, the right side of the
equation is replaced with a constant (8), the mean hourly arrival
rate.

Completed Work Rate=
MIN(Work in Progress/TIME STEP, DELAY3(New Work Rate, time to
completion))
Units:      widgets/Hour
This represents the rate at which work is completed -- assumes some
delay and incomplete mixing of tasks.  The MIN function is used to
ensure that the stock can never go below zero.


Consume Rate=
MIN(Margin/TIME STEP,SMOOTH(Margin/consume time, time to perceive
mismatch on consume))
Units:      buffers/Hour
```

Rate at which resources are consumed during overload to help maintain performance. The MIN function is used to ensure that the stock never falls below zero.

Restore Rate=
SMOOTH((max margin-Margin)/restore time, time to perceive restore opportunity)
Units: buffers/Hour
Represents the rate at which system is able to replenish resources consumed during overload periods. It is subject to a delay, implemented via the SMOOTH function.

Auxiliary variables:

time to completion =
MAX(av time to completion * table for mismatch effect(mismatch)^strength of mismatch effect * effect of consumption on completion(Consume Rate)^strength of consumption effect, 0.5)
Units: Hour
This variable is the current time to complete a unit of work, given the current level of mismatch and consumption of margin. The MAX function ensures time to completion can't fall below 0.5 hours. Time to completion is degraded (increased) by mismatch, but enhanced (decreased) by consuming margin in compensation.

consume time=
MAX(av consume time*effect mismatch on consume(mismatch)^strength mismatch on consume, 0.1)
Units: Hour
Mismatch causes increased consumption of margin resources & buffers, after a threshold

restore time=
MAX(av restore time*effect mismatch on restore(mismatch)^strength mismatch on restore, 0.1)
Units: Hour
Time to restore resources (eg buffers) consumed during overload. MAX function ensures that restoration will take some finite period of time, no matter how much effort is devoted to it.

av consume time = 1
Units: Hour
Baseline time for consumption of margin

av restore time = 2
Units: Hour
Baseline time for restoration of margin

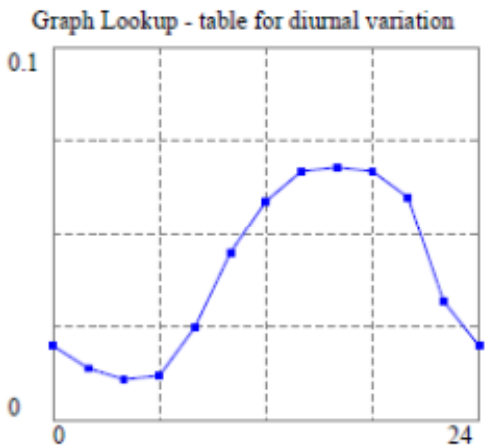
av time to completion = 1.4
Units: Hour
The is the baseline, average time to complete a unit of work under "normal" conditions, ie, not allowing for overload, congestion, etc

Capacity = 15
Units: widgets
Maximum capacity of system before degradation due to overload begins

```
Mismatch =
Work in Progress/capacity
Units:      dimensionless
Ratio of work either being processed or waiting to be processed to
capacity, essentially demand / capacity mismatch

max margin = 50
Units:      buffers
Sets a limit on Margin so it cannot be stockpiled indefinitely or
increase at extraordinarily rapid rate.  Also serves to slow the rate
of restorations as max margin is approached.


time to perceive restore opportunity = 1
Units:      Hour
This variable represents the time required to recognize an
opportunity to rebuild the consumed Margin

time to perceive mismatch on consume = 1
Units:      Hour
Represents the time required to perceive overload (mismatch) and
begin drawing on Margin
```

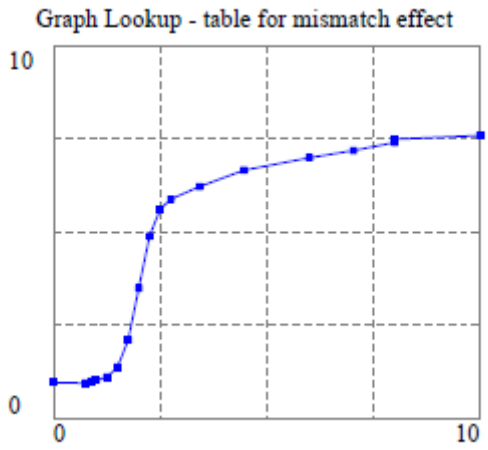
Table variables: These variables create non-linear (typically sigmoid) relations

```
table for diurnal variation([(0,0)- 24, 0.1)], (0,0.02), (2,0.014),
      (4,0.011),(6,0.012),(8,0.025),(10,0.045),(12,0.059),(14,0.067),
      (16,0.068),(18,0.067),(20,0.06),(22,0.032),(24,0.02))
Units:      dimensionless
This variable is used to vary the arrival rate over a 24 hour cycle,
based on typical ED data.  The y-axis values represent the proportion
of total work in a 24 hour cycle that arrives in that 2 hour period.
```

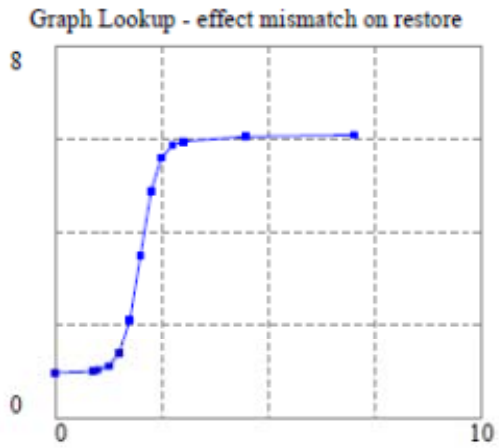


Note that all the following tables produce direct effects on the time to accomplish something: thus an upward sloping graph implies an increase in time, which produces a decrease in the affected rate.

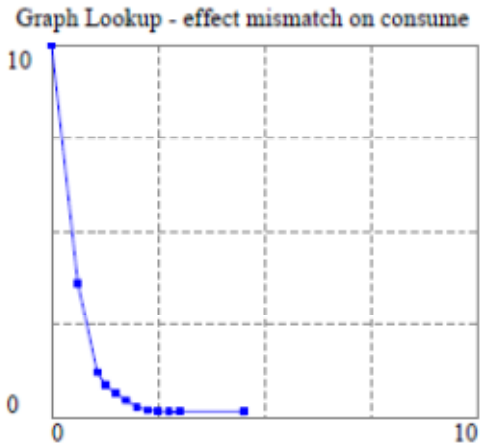
```
table for mismatch effect(
      [(0,0)- (10,10)], (0,0.97), (0.75,0.95), (0.9,1), (1,1.03),
      (1.25,1.11), (1.5,1.38), (1.75,2.11), (2,3.5), (2.25,4.89),
      (2.5,5.62),(2.75,5.89),(3.42508,6.22807),(4.46483,6.66667),(6,7
      ), (7.03364,7.19298),(8,7.41228),(8,7.5),(10,7.6))
Units:      dimensionless
This variable relates the demand / capacity mismatch to an increase
(decrease) in completion time.  Note there is a slight improvement
from low levels of mismatch, to emulate a modest arousal effect from
low degrees of overload.
```



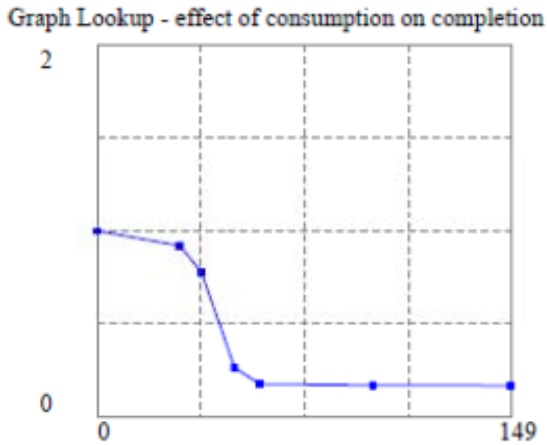
effect mismatch on restore([(0,0)- (10,8)], (0,0.97), (0.9,1), (1,1.03),(1.25,1.11), (1.5,1.38), (1.75,2.11), (2,3.5), (2.25,4.89), (2.5,5.62),(2.75,5.89),(3,5.97),(4.46483,6.07018),(7,6.10526))
Units: dimensionless
This variable causes increasing mismatch to decrease the ability to maintain/restore margin resources, ie, increases restore time



effect mismatch on consume([(0,0)-(10,10)], (0,10), (0.611621,3.64035),(1.07034,1.22807),(1.25,0.901), (1.49847, 0.657895),(1.75,0.474), (2,0.286), (2.25,0.204), (2.5,0.178), (2.75,0.17), (3,0.168),(4.5,0.167))
Units: dimensionless
Variable to relate Mismatch to increases in the consumption of margin resources (reduces time to consume)



```
effect of consumption on completion([(0,0)-(149,2)], (0,1),
(29.6177,0.921053), (37.3639,0.780702), (49.211,0.263158),
(58.3242,0.175439), (99,0.168), (148.5,0.167))
Units:      dimensionless
This variable relates increased consumption of margin resources to
decrease time to completion, thus compensating for mismatch
```



Modifiers for table functions. These variables are used to vary the strength of the effects in the tables above. One will produce the values in the table, 0 will eliminate the tabled effect, < 1 mitigates and > 1 enhances it.

```
strength mismatch on consume = 1
strength mismatch on restore = 1
strength of consumption effect= 1
strength of mismatch effect= 1
```

Simulation Control Parameters

```
FINAL TIME  = 168
Units:      Hour
The final time for the simulation.


INITIAL TIME  = 0
Units:      Hour
The initial time for the simulation.


SAVEPER  = TIME STEP
Units:      Hour [0,?]
The frequency with which output is stored.


TIME STEP  = 0.25
Units:      Hour [0,?]
The time step for the simulation.
```

Experimental control modifications

Pulse challenges were created by adding on or more terms similar to the following to the appropriate rates or accumulations at specified times:

```
height * PULSE(60, 4)
```

Where `height` controls the magnitude of the pulse, 60 is the time of onset and 4 is the duration. A variety of other challenges (pulse trains, ramps, step functions, *etc*) were also explored but the results did not differ qualitatively from those presented in the text.

Appendix 3 Related publications

The following papers related to this project are included for review.

Wears, R. L., Perry, S. J., & McFauls, A. (2006). *Free fall - a case study of resilience, its degradation, and recovery, in an emergency department*. Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France. http://www.resilience-engineering.org/REPapers/Wears_et_al.pdf

Anders, S., Woods, D. D., Wears, R. L., Perry, S. J., & Patterson, E. S. (2006). *Limits on adaptation: modeling resilience and brittleness in hospital emergency departments*. Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Wears, R. L., Perry, S. J., & Nasca, L. (2007). *'Free fall' - highly decentralized, resilient adaptation to demand-capacity mismatches in an emergency department*. Paper presented at the 8th International Naturalistic Decision-Making Conference, Pacific Grove, CA,4 - 6 June 2007.

Wears, R. L., Perry, S. J., Anders, S., & Woods, D. D. (2008). Resilience in the Emergency Department. In E. Hollnagel, C. P. Nemeth & S. W. A. Dekker (Eds.), *Resilience Engineering: Remaining Sensitive to the Possibility of Failure* (pp. 193 - 210). Aldershot, UK: Ashgate.

Wears, R. L., & Perry, S. J. (2008). *A system dynamics representation of resilience*. Paper presented at the 3nd International Symposium on Resilience Engineering, Juan-les-Pins, France. http://www.resilience-engineering.org/REPapers/Wears_et_al.pdf

Gilardi, S., Guglielmetti, C., Perry, S. J., Pravettoni, G., & Wears, R. L. (2009). *Changing horses in mid-stream: sudden changes in plan in dynamic decision problems*. Paper presented at the 9th International Naturalistic Decision-Making Conference, Covent Garden, London, UK,4 - 6 June 2007.

Wears, R. L., & Webb, L. K. (2011). *Fundamental on situational surprise: a case study with implications for resilience*. Paper presented at the 4th International Conference on Resilience Engineering, Sophia Antipolis, France,6 - 8 June 2011.

Morrison, J. B., & Wears, R. L. (2011). *Emergency department crowding: vicious cycles in the ED*. Paper presented at the 29th International System Dynamics Conference, Crystal City, VA,24 - 27 June 2011.

"Free Fall" – A Case Study of Resilience, Its Degradation, and Recovery in an Emergency Department

Robert L Wears[1,2], Shawna J Perry[2], and Allyson McFauls[3]

[1] Clinical Safety Research Unit, Imperial College London, UK
wears@ufl.edu; r.wears@imperial.ac.uk
[2] College of Medicine, University of Florida, USA
sperry@ufl.edu
[3] Department of Emergency Medicine, Shands Jacksonville Medical Center, USA
Allyson.McFauls@jax.ufl.edu

Abstract. Emergency Departments (EDs) are open systems that routinely cope with highly variable and uncertain inputs. This paper will use two critical incidents to explore worker adaptations to complexity and unpredictability, and the organizational interpretation of threats to performance. We use the concept of resilience state space and state transitions to analyse the ED's response to chronic constraints and unexpected shocks.

1 INTRODUCTION

Emergency departments (EDs) are dynamic, open, high-risk systems that function under considerable uncertainty. Like many systems in health care, they have been engineered or designed to only a limited (and some might say naïve) extent. Instead, they have largely evolved sets of artefacts, processes, skills, and attitudes that serve their goals through a process of *bricolage*. These processes support EDs' resilient adaptation to multiple types of variation (*eg*, in numbers of patients, or in the kinds of diagnostic or therapeutic problems encountered), and also to the constraints of economics and human work limits that tend to push them towards working at maximum capacity (Leveson, 2004) and towards the boundary of the safe operating envelope (Cook & Rasmussen, 2005). For the most part, these adaptations are skillfully and unconsciously, almost invisibly performed, as expressed in the Law of Fluency (Woods & Hollnagel, 2006).

However, the resilient capacity of EDs is finite. When it is exceeded, the resulting events offer insight into the ways in which people in the system are sensitive to the possibility of failure; know where to look for evidence of failure and for the resources to cope with it; choose strategies to regain control of the system; and decide which goals to sacrifice in order to meet more important goals and maintain system integrity.

The objective of this paper is to use case studies of two similar events in which the resilient capacity of the ED was exceeded, leaving the system in an uncontrolled state (here called 'free fall'), as a means to explore how resilience is created, lost, and restored in this complex environment.

2 CASE NARRATIVES

Both events occurred in the ED of an inner-city, 653 bed, US teaching hospital that is part of an 8 hospital network. The ED has roughly 90,000 visits per year, and is a Level 1 trauma center. It is subdivided into five major treatment areas totaling 79 beds; two of these areas are dedicated to severe trauma patients and to pediatric cases. Like many US EDs, it experiences severe over-crowding due primarily to a lack of inpatient beds, leading to the 'boarding' of large numbers of admitted patients in the ED (IOM Committee on the Future of Emergency Care in the US, 2006). In response, the ED had reserved one of its non-dedicated treatment areas (comprising 28 beds) for these 'boarders.' One of the remaining two units, with 21 beds, was equipped and staffed for seriously ill patients, and was the site of the episodes described here; the other unit is used only for minor cases. Finally, because the overcrowding problem had previously led to extensive problems with diversion of ambulances en route to EDs in the region, the local public safety authorities had banned the practice of ambulance diversion.

Information on these incidents was gathered by semi-structured interviews of involved staff using the critical incident method, review of documents and personal notes associated with the events, and the ED's volume and through-put records.

2.1 Case 1 – 14 December 2005

At the start of the evening shift (at 1500) on 14 December, the ED was boarding 43 patients. 28 of these filled the unit reserved for boarders, leaving the remaining 15 to be held in a combination of the other two areas and the hallways. Seven were held in the hallway, and all four critical care bays were filled with admitted patients on ventilators. As the shift change rounds in the acute care unit began, the ED received notice that an ambulance was en route with a critically ill patient. Over the course of the next four hours, the ED received by ambulance an additional five critically ill patients (for example, cardiac arrests) requiring ventilator support and other intensive measures, and multiple additional seriously ill but not critical patients (*eg*, chest pain suggestive of heart attack) by ambulance or private conveyance. All treatment spaces were filled; all temporary spaces to hold stretchers were filled; the unit ran out of stretchers and began 'storing' incoming patients in chairs near the nursing station. Congestion was severe, making it physically difficult to move around in the treatment area. This was particularly a problem when new critical patients arrived, since they needed to go to specific treatment spaces because of equipment requirements, and the patients occupying those spaces thus needed to be moved to other locations on very short notice.

The staff later described this situation as a feeling of "free fall", in which they did not know the numbers, types, or problems of the patients in their area of responsibility. The crisis continued until approximately 2200, by which time the staff present felt they had finally gained control of the situation (in the sense of having a clear picture of which patients were present, where they were located, and at least a vague idea of the nature of their problem) and that the system had stabilized.

No identifiable adverse events were associated with this episode, as far as is known.

2.2 Case 2 – 16 November 2005

During the analysis of Case 1, we became aware of a similar incident four weeks prior to it. Events here are structurally almost identical to those outlined in Case 1. The ED was crowded with admitted patients, and the situation had steadily worsened throughout the day. By 1500, there were "… patients everywhere – in chairs, in the aisles. There were no stretchers. We had MICU [critical care] patients from bed 1 to bed 7, and 7 Rescue stretcher patients lined up to be triaged." During this day, the staff recognized that lack of physical space had become the dominant constraint on performance, and attempted a novel adaptation by placing newly triaged, unevaluated cases on stretchers in the hallway. These hallway locations had heretofore only been used for admitted patients for whom no bed was available. Detailed information is available on the trajectory traced by one patient, who suffered an adverse event, as detailed below.

This 58 year old woman presented complaining of severe abdominal pain for several days. She was triaged directly to the hallway since there were no treatment spaces available. The physician performing her initial evaluation was impressed with the seriousness of her condition and felt the problem might require emergent abdominal surgery. She switched this patient with another of her own patients in a routine treatment area, in order to have enough privacy to do a proper physical examination (including a pelvic exam), and then moved them back to their original locations. The routine investigations for an acute abdomen case were ordered, including a plain film (x-ray) of the abdomen. Twice, the patient was moved to x-ray but had to return without radiography because all the technicians were busy with cases in the trauma unit. Finally, near the shift change at 2300, a decision was made to order a computed tomography (CT) scan of the abdomen in anticipation of eventually getting a negative result from the plain film. This decision was influenced by several factors: 1) desire to have a "clear plan" for the oncoming shift; 2) knowledge that the surgical team was similarly overwhelmed in the trauma unit and would be unable to break away to evaluate the patient for some time; 3) the general opinion that the plain film rarely adds important information in these cases; 4) knowledge that the surgeons would probably request the CT prior to their evaluation to save time; and 5) knowledge that an abdominal CT often takes several hours to complete. Eventually, the plain film was obtained, but due to the congestion and confusion in the area, and the discussion about the CT scan at shift turnover, it was not read prior to the administration of oral contrast material in preparation for the CT. Unfortunately, the plain film showed free air, indicating the perforation of a hollow organ (such as the stomach or intestine). In a perforation, oral radiographic contrast material is contraindicated because it can spill out through the perforation and cause a chemical peritonitis, aggravating an already severe condition and complicating the required surgery. The radiologist eventually read the plain film before the CT was performed but after the patient had been given oral contrast, and alerted the ED to the problem. The surgeons were then called, and the patient was taken to the operating room where a perforated ulcer with extensive peritonitis was repaired successfully. Post-operatively the patient suffered a severe stroke; the relationship of this to the preceding events is unclear.

Ironically, a meeting to investigate the cause of this patient's injury was held on 14 December (the date of Case 1), prompting one of the participants to remark, "… the same thing is happening out there again today."

3 ANALYSIS

These cases represent episodes where the resources and coping strategies that normally provide resilience against variation and the unexpected became exhausted, and workers had to adopt new strategies and make sacrifice decisions, abandoning lower level goals in order to preserve higher ones and regain control of the situation (Cook & Nemeth, 2006).

State space model. The resilience state space model (Hollnagel & Sundström, 2006) provides a compact way to summarize the state of the ED as it progressed into and out of this crisis (see Figure 1). Both shifts began in the state of 'regular reduced functioning' – this was not 'normal functioning', at least in the normative sense, although this was certainly the most common state of the ED at this season. The 'normalization of deviance' (Vaughan, 1996) had insidiously consumed the ED's buffering capacity, such that the capability to absorb sudden disruptions had been degraded. This represents a state of chronic decompensation in the system; it remained operable, but at a reduced level of functioning and a reduced margin of safety. Because of the loss of buffering ability, the ED was more tightly coupled to the inpatient beds than was normally expected.

As a result, when the number of critical and serious patients needing assessment and intervention grew rapidly, (and seemingly without limit), the ED shifted to 'irregular reduced functioning'. This was marked by an attempt to continue with diagnostic and therapeutic measures in all patients, using irregular spaces and informally supported sacrifices of some routine procedures. In Case 2, the novel adaptation of triaging newly arrived patients to the hallway when stretcher spaces were exhausted, is an example of this strategy of trying to use novel spaces to maintain some reduced level of functioning. Essentially this was a strategy to develop new compensatory buffers to help manage the disturbance.

One interesting aspect of these adaptations was the strategy of placing patients in chairs. It was never spoken explicitly, but widely recognized, that the ability to maintain postural tone (*ie*, to sit in a chair) was an indicator of a certain level of stability; thus management of patients in chairs could be sacrificed in order to attend to patients of higher criticality. In effect, this strategy identifies patients who might be physiologically more resilient, and "borrows" some of their resilience to provide additional buffering capacity to support higher level goals and operations.

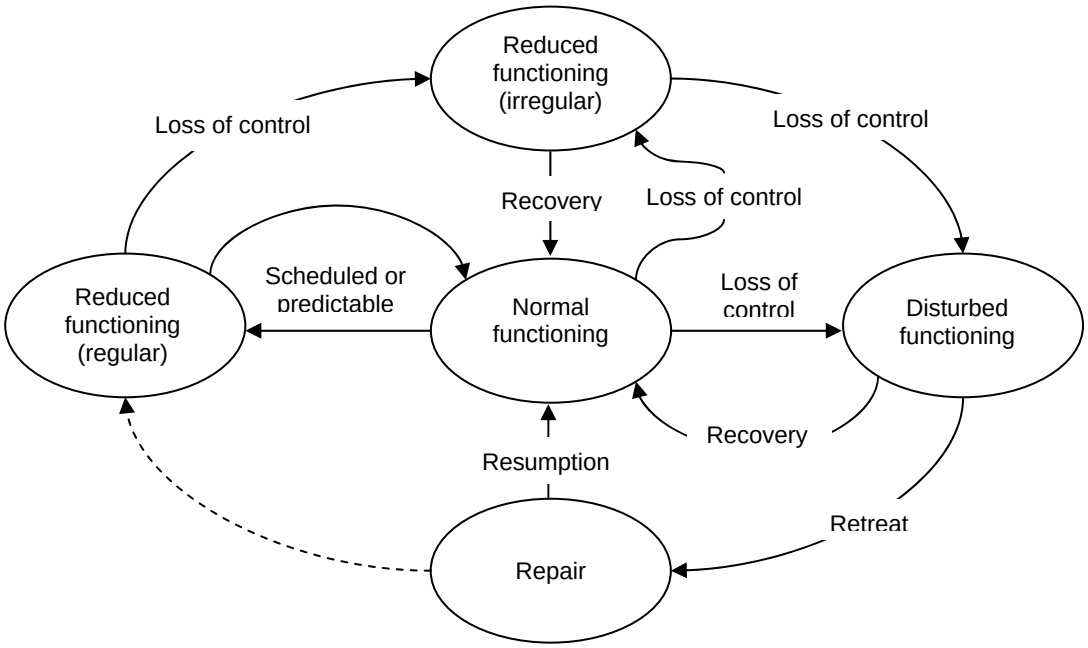


Fig. 1. State-space diagram for service organizations. (After Hollnagel & Sunström 2006, p 341, used with permission)

A second strategy involved sacrificing some lower level goals in order to be able to satisfy higher one. An illustration of this behaviour can be found in the timing of electrocardiograms (ECGs) for chest pain patients. A national standard has been proposed that any chest pain patient should receive an ECG within 15 minutes of arrival to the ED. In this ED, due to chronic decompensation, the mean time to ECG is typically around 35 minutes; in Case 1, the mean increased to 52 minutes (range 0 to 154 minutes).

A third adaptive strategy seen at this level, in both cases, was an anticipatory attempt to use 'feed-forward' techniques to facilitate routine operations in the future. This strategy assumes that the current disturbance will be transient, so the goal should be facilitating those functions that will be important on resumption of more nearly normal operations. In both cases, physicians used a strategy of anticipatory test ordering to try to 'save time in the future'. Here, instead of selecting tests in series, specifically tailored to a patient's condition (which would require a detailed assessment for which there was no time), physicians would order a broad battery of tests in parallel, assuming that by the time the results came back (typically in several hours), they would have completed that detailed assessment and would thus know which results were not relevant. This offers obvious advantages over waiting to place the order, since the results would then be even further delayed. This can be viewed also as a strategy for shifting some of the overload to other parts of the organization, is a mechanism by which the disturbance spreads; it also tightness the coupling between the lab and the ED. In Case 2, this strategy led to placing the order for the CT scan without first reviewing the plain film and other, simpler tests.

In both cases, the situation eventually worsened to 'disturbed functioning', where additional and highly irregular resources were employed. For example, a small office for

the attending physician adjacent to the treatment area was used to perform ECGs on patients who were waiting in the aisles or in chairs, since it had a door that could be closed for privacy. Similarly, a small closet normally used for storage of respiratory and advanced airway equipment was used as a blood drawing area.

Finally, the ED was forced to retreat entirely from any semblance of routine operations for any but the most time-critical of patients. Essentially, this was a strategic decision to concentrate on 'disturbance management', and was manifested by a shift in operations from medical content to simple tracking – identifying patients, the (irregular) spaces to which they were assigned, and a vague categorization of problem type. In both cases, this was aided by creating a second status board within the ED's main status board. This second board was used for patients without assigned treatment areas who were waiting in chairs around the nursing stations, and listed only patient's name, location (this required some informal inventions, *eg*, 'Pyxis chair 2') and check boxes indicating that a physician had spoken to them, and that blood had been drawn. This is essentially the 'repair' state, and can be viewed as a strategy to stop continuing operations in an attempt to regain control. In terms of goal states, it involves the sacrificing of most lower and intermediate level goals in order to preserve resources to restart the system once the disturbance had passed. (It is undoubtedly not accidental that this strategy is expressed in the rhetoric of defeat and resignation).

Once the repair had been successfully accomplished (in that workers now knew which patients they had responsibility for, where those patients were physically located, and what their basic problem type was), and the system stabilized (aided by the decrease in the numbers of incoming critical patients), then normal operations could be gradually resumed. This was done cautiously; it took some time to build up confidence that the current assessments were accurate and complete – the "continuing expectation of future surprise" (Rochlin, 1999) led to a conservative and gradual re-starting of routine operations.

The rapidity of the degradation in performance suggests that the ED possesses highly nonlinear characteristics. The flow of patients through the department on these days seems analogous to phase shifts in the state of matter; discontinuous transitions from laminar, to turbulent flow, to complete stagnation, similar to the condensation of water from a vapor, to a liquid, to ice.

Other adaptations. Other adaptations also played a role in the recovery, albeit to a more limited extent. In Case 1, the crisis became apparent during normal working hours, so additional attending physicians were available to come to the ED to assist. These additional staff were helpful, but were hobbled by the general congestion (in fact, they added a bit to it). Similarly, the hospital's nursing supervisor on duty in Case 1 was widely thought to be one of the more effective, and her presence during the episode assisted in temporarily shifting some ventilator patients to non-standard areas (such as the trauma receiving unit) to regain valuable treatment space, and in caring for incoming critical cases.

Organisational response. In contrast to the worker adaptations performed dynamically in context, the higher level organizational responses to these events were delayed and muted. In Case 2, the specific adverse event was reviewed by an internal quality group, but the span of control of this group was limited, so no general review of the mismatch between resources and demand occurred; instead, the issue was referred upward to hospital administration, where it languished.

In Case 1, in part because no patient was apparently harmed no 'after action review' was held to analyze the hazards or vulnerabilities underlying the episode, despite requests from involved staff.

DISCUSSION

These cases illustrate a complex pattern of performance degradations: acute decompensation, superimposed on chronic decompensation (Miller & Xiao, 2006). The ability of the staff to compensate during the period of chronic decompensation masked the drift toward the boundary of failure. This proximity to failure was finally revealed when buffers that were not easily further expanded were exceeded. Specifically, the lack of available physical space became the irreducible constraint in both cases that led the system ultimately to transition to the repair state.

Clinicians who self-select to work in EDs have a high tolerance for uncertainty, and take great pride in their ability to respond resiliently to uncertain and unpredictable demands. The informal motto: "Anyone, anything, anytime", which was used as the title for a recent history of emergency medicine (Zink, 2006), neatly expresses this common ethos. In terms of patient load, the demands in both these cases were not extraordinary; the total daily visits on these days were close to the ED's average volume, and the acute care unit had successfully managed mass casualty incidents – large numbers of critically ill patients arriving simultaneously or in rapid succession – on numerous occasions in the past. Therefore, the sensation of "free fall" experienced on these two days was highly distressing to the health professionals involved. Rather than being able to "take things in one's stride", as they normally expect to do, they were confronted with an acute sense of overwhelming failure and lack of control (Cook & Nemeth, 2006). Although they did not have the language of the resilience state space in which to express it, the distress that many senior, experienced workers felt over these incidents likely stems from this being their first, ever, transition into the repair state. Since by definition, an ED should never be in the repair state, such a transition challenges the very core of their collective professional identity. In addition, the impression that these episodes were related to hospital management issues, rather than external events (such as a hurricane or other disaster), added a sense of abandonment, which increased the affective impact on the workers.

Resilience in this setting is dynamic and adaptive, but finite in capacity. Three characteristic shifts in strategy accompany changes in the 'resilience state' of the system. These strategies are: attempts to increase buffering capacity; sacrificing lower level goals to preserve higher; and using feed forward methods to facilitate future functional-

ity in anticipation of returning to normal operations. These adaptive strategies are generally, but not always successful, and sometimes bear risks of their own. However, their net effect seems to be to move the system from unstable to stable conditions and to allow the resumption of normal operations.

REFERENCES

Cook, R. I., & Nemeth, C. (2006). Taking things in one's stride: cognitive features of two resilient performances. In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 205 - 221). Aldershot, UK: Ashgate.

Cook, R. I., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. *Quality & Safety in Health Care, 14*(2), 130-134.

Hollnagel, E., & Sundström, G. (2006). States of resilience. In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 339 - 346). Aldershot, UK: Ashgate.

IOM Committee on the Future of Emergency Care in the US. (2006). *Hospital-Based Emergency Care: At the Breaking Point*. Washington, DC: National Academies Press.

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science, 42*(4), 237 - 270.

Miller, A., & Xiao, Y. (2006). Multi-level strategies to achieve resilience for an organisation operating at capacity: a case study at a trauma centre. *Cognition, Technology & Work*, 1-16.

Rochlin, G. I. (1999). Safe operation as a social construct. *Ergonomics, 42*(11), 1549-1560.

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*. Chicago, IL: University of Chicago Press.

Woods, D. D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL: CRC Press / Taylor & Francis Group.

Zink, B. J. (2006). *Anyone, Anything, Anytime: A History of Emergency Medicine*. Amsterdam, NL: Elsevier.

Limits on Adaptation: Modeling Resilience and Brittleness in Hospital Emergency Departments

Shilo Anders[1], David D. Woods[1],
Robert L. Wears[2, 3], Shawna J. Perry[3]
Emily Patterson[1]

[1] The Ohio State University, US
anders.41@osu.edu, woods.2@osu.edu
patterson.150@osu.edu
[2] Imperial College, London
wears@ufl.edu; r.wears@imperial.ac.uk
[3] University of Florida, US
sperry@ufl.edu

Abstract. The emergency department is a complex, highly adaptive system that operates in the face of uncertainty and limited resources. Field observations of an emergency department were conducted to investigate properties of resilience and adaptive challenge. A specific case was explored in order to make generalizations about the classes of adaptive challenge. In addition, researchers used this case to illustrate how the emergency department adapts as load increases in terms of the five properties of resilience in action that is grounded in actual observations.

1 INTRODUCTION

Some systems are designed to adapt to changing demands such as a hospital's emergency department. The emergency department is as a complex, dynamic setting where successful and effective work must occur in the face of high consequences of failure, practitioners are operating under time and resource pressures, and competing goal conflicts.

Analyzing how examples of such systems are adapted to potentially changing demands and studying how they adapt as load increases can reveal a great deal about how to design resilient organizations. Ironically, hospital emergency departments are also critical pressure points in the U.S health care system. In spite of be adaptive by design, recent assessments see emergency departments as a highly brittle component of the overall healthcare system (see Committee on the Future of Emergency Care in the US, 2006). Emergency departments are under resource pressures and face new demands which can lead to coordination breakdowns at boundary conditions (e.g. overcrowding, lack of coordination and boarding patients for other units).

This paper analyzes an active emergency department in terms of resilience concepts, in particular, to test the Westrum taxonomy of resilience situations and further refine the properties of resilience (Westrum, 2006; Woods, 2006). The data are based on observation of an emergency department as it handles different loads and retrospective analyses of actual cases of situations that drove this system very near its limit in adaptive capacity requiring a shift from one level in the taxonomy to another.

The strategies for adaptation are organized around four classes of adaptive challenge. A routine day is one in which the system is operating under usual conditions and described by practitioners as "run of the mill" where the system anticipates shifts beyond the routine and adapt apparently seamlessly. In a second class of situations, a key person recognizes system degradation as load and demands are increasing, thus initiates adaptive tactics (e.g., recruiting and reorganizing multiple resources) to manage the challenges and maintain performance. In other situations the demands increase to the point that the needed adaptations occur at the level of the whole department. In the latter two classes, the demands on the organization challenge its ability to sustain operations and risk escalating to a breaking point, which has been described by practitioners as a "free fall" (e.g., Wears, Perry, & McFauls, 2007). Practitioners have to recognize and anticipate the trend and to reorganize activities and resources at the same time as they are struggling to handle patient load, The last class of situations are planned for but rarely experienced events that call for a complete planned reorganization in the wake of a catastrophic event, e.g. a mass casualty event (Perry, Wears, & Anderson, 2007).

2 METHOD

Observations for this paper were done in a single emergency department during a four-day period and included brief follow-up interviews with the attending physician after the observed shift. A specific case that illustrates how medical personnel cope with complexity and illustrates a transition from a "run of the mill" day to a second class of adaptive challenge is analyzed using a process tracing technique.

3 CASE STUDY

3.1 The Setting

The emergency department in the observed hospital consists of four areas: medical, trauma, pediatric, and flex-care (least critical patients). The observer spent the most time in the medical and trauma units, which is were the specific incident occurred and will be discussed in detail. The trauma unit is actually connected to the medical unit via a doorway leading into the area of the most critical patient beds in the medical (see Figure 1).

The medical unit consists of four critical beds (e.g. patients that need to be on a ventilator), 15 other beds loosely descending in order of criticality, and a "fishbowl" where a sitter is present for the psychiatric patients. The staffing consists of a shared emergency department attending that also manages the trauma unit, four residents (a chief of the day, two other residents, and one that manages both the critical medical beds and trauma), five nurses, nurse of the day, charge nurse (responsible for all emergency department units), and two technicians.

The trauma unit consists of five beds and is meant to be a resuscitation area where patients are stabilized before being moved to other areas in the hospital. The staffing consists of two nurses, one medical technician, the shared resident, an on-call surgical attending, and the shared attending.

3.2 Case Study

Before the escalating event occurred in the emergency department, the night seemed to progressing in what could be described a "run of the mill" shift. The attending spends time shifting patients and deciding where to send the less critical patients in order to free up space in the units. Throughout the evening a steady flow of patients, in both units under observation. The medical unit has only one critical bed occupied, while the trauma unit actually received a number of patients earlier as well as from the night before, hence it only had one open bed. The patients were all stable and personnel were waiting to transfer these patients to other areas of the hospital. Of these patients two were on ventilators, while the other two were conscious. This is the setting for the following case which is described in a linear fashion with commentary from researchers about the properties of resilience.

Table 1. Case description and comments

Case	**Commentary**
The trauma unit of the emergency department receives a call about 3 incoming patients. In order to accommodate these patients, one current patient is admitted to the hospital, and another is moved to the hall.	The unit can only handle one more patient without reconfiguring. Therefore, they are too close to the **margin** if all 3 anticipated patients arrive given the current **capacity**. They reconfigure by moving a patient to the hospital and moving one patient to an area where ventilators cannot be used.
Patient 1 (first expected of 3 patients) comes into the trauma center and is very combative due to head trauma, so before he can be sedated, he is physically restrained by about 8 people.	By using the relatively large resource of eight people now to sedate the patient, he will require less active monitoring later.
Two more patients arrive. The first is the second expected patient of three. The second is her child, who was not expected. The first is put in the open bed, while the child is taken to the pediatric emergency department. The pediatric fellow who transferred the child had recently arrived to assist with the new patients in response to a standard page given to all physicians when critical patients are due to arrive, but had not been aware that a pediatric patient was expected.	In order to make observable all new critical patients to the emergency department all attendings and fellows are paged for any critical patients. When an unexpected child arrives rather than helping in the trauma unit the pediatric fellow changes plans taking the child to the emergency pediatric unit herself.
The unit is alerted that the third expected critical patient should arrive in less than 5 minutes. The	**Buffering capacity** is increased by creating more beds before they are

attending asks the observer to get the chief resident from the medical emergency department to help. The least critical of the patients is wheeled into the hallway (next to the 2 patients already in the hall). The first patient is intubated and second patient is assessed.	needed. In addition, it is a better buffer in that it allows the use of ventilators, which is not possible in the hallway.
Patient 4 arrives from an unrelated accident. The charge nurse asks the paramedic to page the nurse manager to get additional nursing staff. This patient is intubated at the same time as patient 1. The surgical attending arrives to decide which patient should be operated on first.	The charge nurse realizes that the trauma unit's resources (nursing staff) is running out. She unsuccessfully attempts to access resources from a larger resource pool (nursing for the entire hospital) as a **cross-scale interaction** attempt to find additional resources in order to increase the distance between the current state of the system and the safety boundary. The surgical attending is opportunistically deciding which patients would benefit most from surgery, which also frees up trauma resources.
The attending asks the radiology resident that is in the trauma unit to carefully examine all of the x-rays and report any abnormal findings to the trauma attending in order to minimize missing anomalies.	Attending realizes that in this state it is likely that an important alert might be missed, so she recruits other resources as a checking mechanism.
Patient 5 (husband of mother and child from car accident) arrives. All of the beds are taken and no more patients can be put into the hallway without blocking access. The attending asks the trauma charge nurse which patient is most stable and could be moved to a medical ED bed.	The trauma unit is reaching a **boundary** in that it has no more resources available within the unit itself, so in order to avoid collapse, the system shifts to utilization of resources in the medical unit.
Patient 6 arrives with a knife wound. He is quickly examined and the charge nurse has the paramedics wait with the patient on the stretcher in the corner of the room until they have time to process him.	Personnel from outside the emergency department are recruited to monitor the patient in a holding pattern.
Patient 1 is taken to CT scan, and patient 5 is moved from the stretcher to a bed.	One resource reduction strategy employed at several points is reducing patient movement by doing tasks in the emergency department,

	but CT scans are not able to be moved due to the heavy equipment.
The unit receives word that another critical patient (7) is en route. The trauma charge nurse tells the medical charge nurse to expect a patient. Another pediatric patient who had previously been moved to the hallway to make room for the other patients is moved to the pediatric unit to make more room in the hallway.	Bed and staff resources are flexibly recruited from other units, including the medical and pediatric unit. This recruitment signals an understanding that the situation is **precarious** in the sense that they are near the edge of what they can tolerate with current resources.
The new critical patient 7 (hip fracture from car accident) arrives before a bed is made available, so ends up taking the space of the patient getting a CT scan. Patient 5 is prepared for a chest tube. In all, 24 caregivers are in a small, noisy space, primarily caring for patients 5, 6, and 7.	Although there are many patients, most resources are dedicated to a small number of prioritized patients.
The medical charge nurse starts triage and intake of patient 8 (intoxicated patient who had driven into a telephone pole) in the hallway. Another nurse from the medical ED assists the trauma nurses with patient care.	Facilitation occurs **flexibly** by sharing resources across the trauma and medical units.
Two more patients (knife wound and bleeding from artery due to an accidental wound) walk into the trauma unit. A medical ED bed is designated for on-site treatment by two resident physicians from the operating room. Three patients with minor wounds are stitched sequentially. Patient treatment continues without further incident for all other patients.	**Buffering capacity** in the operating room is increased given anticipated needs of critical patients by a non-routine strategy to provide surgical care.

4 DISCUSSION

The data about how an emergency department adapts as load increases provide the means to investigate the five properties of resilience in action in a realistic organization (Woods, 2006):

- the size or kinds of disruptions the system can absorb or adapt to without a fundamental breakdown in performance or in the system's structure (buffering capacity);
- the system's ability to restructure itself in response to external changes or pressures (flexibility versus stiffness);
- where the system is currently operating relative to one or another kind of performance boundary (margin);

- how a system behaves near a boundary – whether the system gracefully degrades as stress/pressure increase or collapses quickly when pressure exceeds adaptive capacity (tolerance); and
- cross-scale interactions, both upward—as when the ED makes demands on the larger hospital system to adapt to high load and downward—as when the hospital/care system adapts in ways that restrict the adaptive capacity of the ED.

Each of this ideas will be further elaborated on in the context of the emergency department case describe above.

The buffering capacity of the emergency department is dynamically generated and increased throughout this incident when the medical personnel recognize that their resources are depleted and the margin in reaching a breaking point. During this incident the attending recognizes that trauma unit in isolation can no longer provide adequate patient care, so she reconfigures the system by pulling resources from the medical unit. This reconfiguration is lead by a key figure, the attending, which cascades to others.

The notion of the buffer size changes as the scenario unfolds, such that initially the trauma unit accommodates the new influx of patients by "creating" beds in the hallway, but this strategy turns out to not be adequate to handle the patient load, so further adaptation must occur. The capacity of the trauma unit was smaller than the actual patient need, such that external resources had to be utilized in order to prevent collapse (see Fig 1 for illustration of where buffering capacity is increased). The trauma unit utilized these back-up resources that are in the margin zone to create resilience rather than undergoing a re-organization (Miller & Xiao, 2006).

Specifically the hallway became a patient holding area, the pediatric unit took extra patients, and similarly the medical unit took extra patients as well as had one bed turn into a "mini" operating room. Monitoring of the resources was more static and made observable to the distant units of the emergency department via the paging system. Hence, the trauma unit was able to off-load the pediatric patients that were taking up needed resources in the trauma unit. Additionally, how and when these additional resources are deployed depend on a variety of factors. This include where the system is in terms of its perceived distance from the margin and availability and timeliness of resources.

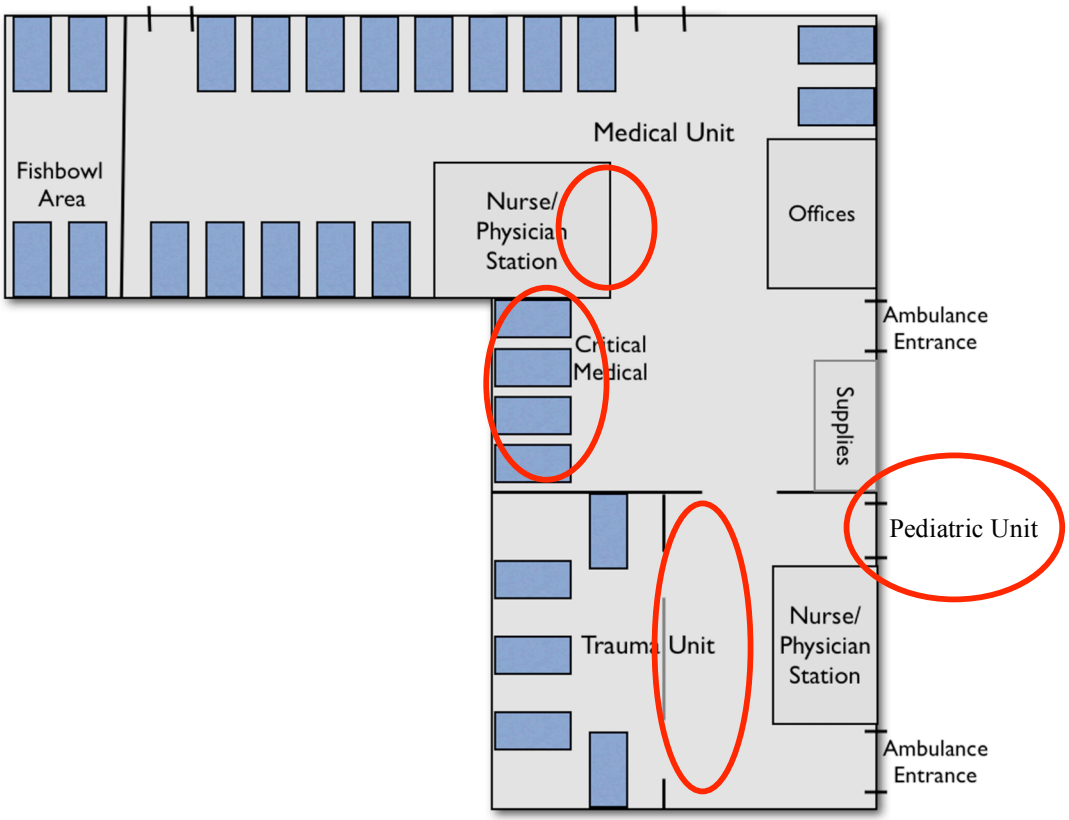


Fig 1: Layout of the medical and trauma units of the emergency department overlaid with areas utilized for patient care that were outside the normal system functions

In order to address the challenges of this case, available resources performed functions outside the scope of normal practice. The flexibility required to do this is a property of resilience, without this flexibility the system would fail. The precariousness of the trauma unit is realized when the necessary resources are no longer available. As is illustrated in Figure 2, initially the resources of the trauma unit are able to cope with the situation, but as the situation escalates performance and resources degrade. The emergency department compensates by utilizing resources from the other units (medical and pediatric). In order to do that the attending made a sacrifice decision to abandon the goal of using resources only from the trauma unit, thus keeping other units free for other potential emergencies to using other units in order to maintain control of the situation (Cook & Nemeth, 2006).

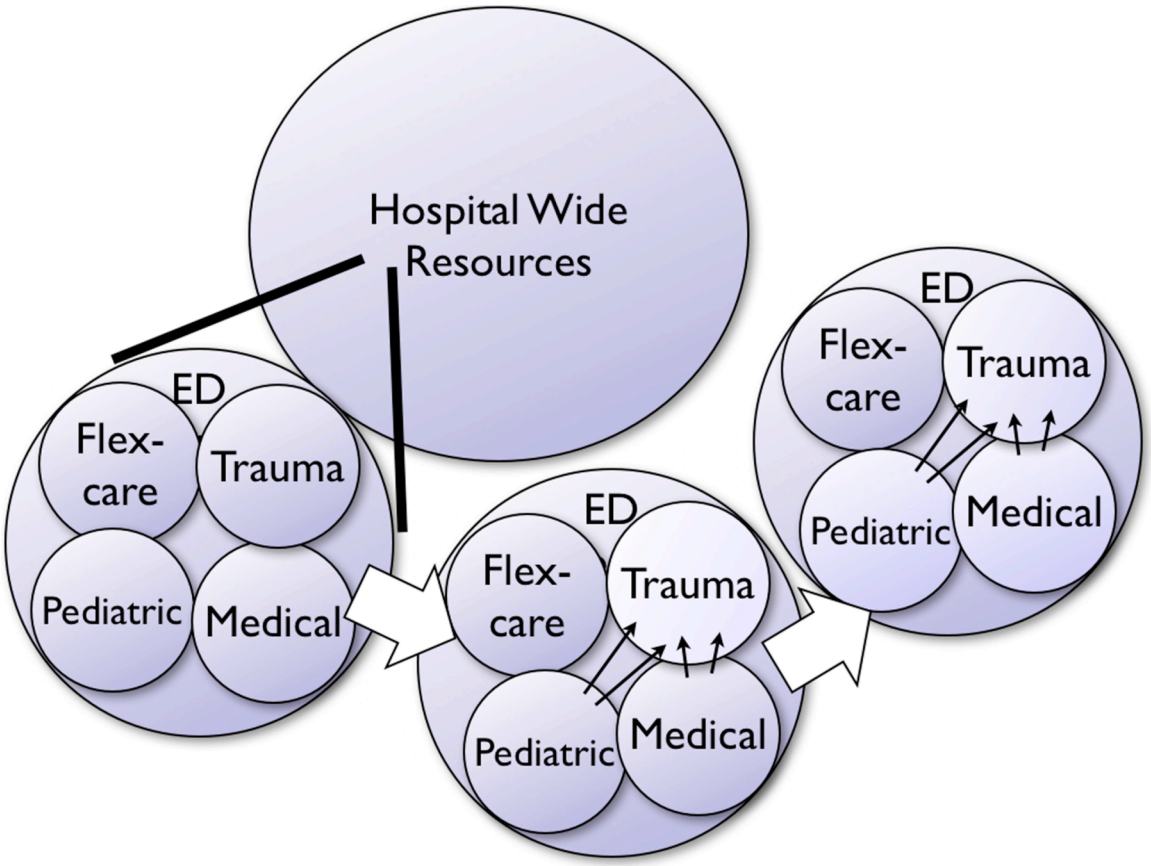


Fig 2: Representation of resource allocation within the hospital and where resources were garnered

Finally, not only does the trauma unit coordinate within the emergency department, but also with resources outside it. The paramedics take over patient care for a less critical patient while the other personnel attend to the more critical patients.

5 CONCLUSIONS

The emergency department exhibits properties of resilience in the way patient care is coordinated. In the current paper a single case was examined in terms of the five properties of resilience, which create a framework for classes of adaptive challenge. In maintaining a balance of these properties potential for collapse can be perceived and adapted for in advance, thus changing the class of adaptive challenge.

REFERENCES

Cook, R. I., & Nemeth, C. (2006). Taking things in one's stride: Cognitive features of two resilient performances. . In Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.), *Resilience Engineering* (pp. 205-220). Aldershott, UK: Ashgate.

Committee on the Future of Emergency Care in the US. (2006). *Hospital-based emergency care: At the breaking point.* Washington, D.C.: National Academies Press.

Miller, A. & Xiao, Y. (2006). Multi-level strategies to achieve resilience for an

organization operating at capacity: A case study at a trauma centre. *Cognition, Technology, Work, 8*.

Perry, S. J., Wears, R. L., & Anderson, B. (2007). Extemporaneous adaptation to evolving complexity: A case study of resilience in a healthcare setting. Submitted *2nd Symposium on Resilience Engineering,* Cannes, France.

Wears, R. L., & Perry, S. J. (2007). "Free fall" – A case study of resilience, its degradation, and recovery in an emergency department. . Submitted *2nd Symposium on Resilience Engineering,* Cannes, France.

Westrum, R. (2006). A typology of resilience situations. In Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.), *Resilience Engineering* (pp. 55-67). Aldershott, UK: Ashgate.

Woods, D. (2006). Essential characteristics of resilience. In Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.), *Resilience Engineering* (pp. 55-67). Aldershott, UK: Ashgate.

'Free Fall' – Highly Decentralized, Resilient Adaptation to Circumstances beyond Operational Experience

Robert L Wears, MD, MS
University of Florida / Imperial College London
wears@ufl.edu

Shawna J Perry, MD
University of Florida
sperry@ufl.edu

Leonardo Nasca, MD
University of Florida
leonardo.nasca@jax.ufl.edu

ABSTRACT

Emergency departments are open systems that routinely cope with highly variable and uncertain inputs under constraints of resources and time. This paper uses two critical incidents to explore worker adaptations to complexity and unpredictability in operational circumstances beyond the bounds of the "normal, natural troubles" with which they are familiar. We use the concept of resilience state space and transitions among states to analyze the emergency department as a joint cognitive system that dynamically responds to unexpected shocks and chronic constraints that carry operations beyond the bounds of experience.

Keywords

Resilience, adaptation, health care, emergency services.

INTRODUCTION

Emergency departments (EDs) are dynamic, open, high-risk systems that operate under considerable uncertainty. Like many systems in health care, they have been engineered or designed to only a limited (some might say naïve) extent. Instead, they have largely evolved sets of operational artefacts, processes, skills, and attitudes through a sort of *bricolage*. These processes support EDs' resilient adaptation to multiple types of variation (*eg*, in numbers of patients, or in the kinds of diagnostic or therapeutic problems encountered), and also to the constraints of economics and human work limits that tend to push them towards working at maximum capacity (Leveson, 2004) and towards the boundary of the safe operating envelope (Cook & Rasmussen, 2005). Many of these adaptations serve as readily available solutions to the "expected, normal and natural troubles" with which workers have become familiar through experience, and word-of-mouth (Voß, Procter, Slack, Hartswood, & Rouncefield, 2006). For the most part, these adaptations are skillfully and unconsciously, almost invisibly performed, as expressed in the Law of Fluency (Woods & Hollnagel, 2006). They are the usual solutions to the usual problems, and thus are contained with a *horizon of tractability* (Voß *et al.*, 2006).

However, the resilient capacity of EDs is finite. Its limits are not commonly reached – if they were, the organization would cease to exist – so they are typically only exceeded by circumstances beyond operational experience. In such cases, the resulting events offer insight into the ways in which people in the system are sensitive to the possibility of failure; know where to look for evidence of failure and for the resources to cope with it; choose strategies to regain control of the system; and decide which goals to sacrifice in order to meet more important goals and maintain system integrity.

The objective of this paper is to use analyses of two similar events in which the resilient capacity of the ED was exceeded, leaving the system in an uncontrolled state (called 'free fall' by the subjects experiencing it), as a means to explore how workers adapt independently but in a distributed, coordinated fashion to threats to the operational integrity of the system.

NARRATIVE DESCRIPTIONS

We have previously described these events in more detail from the point of view of resilience engineering (Wears & Perry, 2006). We briefly recapitulate those descriptions here. Both events involved a gradual overwhelming of the ED's capacity to function by the infortuitous conjunction of an influx of critical patients and the loss of physical space to store patients due to hospital overcrowding.

Setting

Both events occurred in the ED of an inner-city, 653 bed, US teaching hospital. The ED has roughly 90,000 visits per year, and is a Level 1 trauma center. It is subdivided into five major treatment areas totaling 79 beds; 2 of these areas are dedicated to severe trauma patients and to pediatric cases, respectively. Like many US EDs, it has for several years experienced severe over-crowding due to a lack of inpatient beds, leading to the 'boarding' of large numbers of admitted patients in the ED (IOM Committee on the Future of Emergency Care in the US, 2006). In response, about 1 year prior to these events, the ED reserved one of its non-dedicated treatment areas (comprising 28 beds) exclusively for these 'boarders.' One of the remaining two units, with 21 beds, was equipped and staffed for seriously ill patients, and was the site of the episodes described here; the other unit is used only for minor cases. Because the overcrowding problem had previously led to extensive problems with diversion of ambulances en route to EDs in the region, the local public safety authorities had banned the practice of ambulance diversion.

Information on these incidents was gathered by semi-structured interviews of involved staff using the critical incident method, review of documents and personal notes associated with the events, and the ED's volume and through-put records.

Case 1

At the start of the evening shift (1500), the ED was boarding 43 patients; 28 of these filled the unit reserved for boarders, the remaining 15 were split among the other two areas and the hallway separating the units. Seven were held in the hallway; all four of the acute care unit's critical care bays were filled with admitted patients on ventilators. As the shift change rounds began, the ED received a critically ill ambulance patient. Over the course of the next four hours, an additional five critically ill patients requiring ventilator support and other intensive measures arrived, in addition to multiple additional seriously but not critically ill patients (*eg*, chest pain suggestive of heart attack). All treatment spaces were filled; all temporary spaces to hold stretchers were filled; the unit ran out of stretchers and began 'storing' incoming patients in chairs near the nursing station. Congestion was severe, making it physically difficult to move around in the treatment area. This was particularly a problem when new critical patients arrived, since they needed to go to specific treatment spaces because of equipment requirements, and the patients occupying those spaces thus needed to be moved to other locations on very short notice.

The staff later described this situation as a feeling of "free fall", in which they did not know the numbers, types, or problems of the patients in their unit. The crisis continued until approximately 2200, by which time the staff felt they had finally gained control of the situation (in the sense of having a clear picture of which patients were present, where they were located, and at least a vague idea of the nature of their problem) and that the system had stabilized. No adverse events were associated with this episode, as far as is known.

Case 2

On a morning 4 weeks prior to Case 1, the ED was again crowded with admitted patients; the situation steadily worsened throughout the day. By 1500, there were "… patients everywhere – in chairs, in the aisles. There were no stretchers. We had [critical care] patients from bed 1 to bed 7, and 7 [ambulance] stretcher patients lined up to be triaged." During this time, the staff recognized that lack of physical space had become the dominant constraint on performance, and attempted a novel adaptation by placing newly triaged, unevaluated cases on stretchers in the hallway. These hallway locations had heretofore only been used for admitted patients for whom no bed was available. Detailed information was available on the trajectory of one patient who suffered an adverse event, described below.

The patient was a 58 year old woman complaining of severe abdominal pain for several days, who was triaged directly to the hallway because no treatment spaces were available. The ED physician performing her initial evaluation felt her condition was serious and might require emergent abdominal surgery. She switched this patient with another of her own patients in a routine treatment area, in order to have enough privacy to do a proper physical examination (including a pelvic exam), and then moved them back to their original locations. The usual investigations for an acute abdomen case were ordered, including a plain film (x-ray) of the abdomen. Twice, the patient was moved to x-ray but returned without radiography

because the technicians were busy with cases in the trauma unit. Finally, near the 2300 shift change, a decision was made to order a computed tomography (CT) scan of the abdomen in anticipation of eventually getting a negative result from the plain film. This 'feed-forward' decision was influenced by several factors: 1) desire to have a "clear plan" for the oncoming shift; 2) knowledge that the surgical team was busy in the trauma unit and would be unable to evaluate the patient for some time; 3) the general opinion that plain films rarely add important information in these cases; 4) the expectation that the surgeons would request the CT prior to their evaluation to save time; and 5) knowledge that an abdominal CT often takes several hours to complete. Eventually, the plain film was obtained, but was not read prior to the administration of oral contrast material in preparation for the CT. Unfortunately, the plain film showed free air, indicating the perforation of a hollow organ (such as the stomach or intestine). In a perforation, oral contrast material is contraindicated because it can spill out through the perforation and cause a chemical peritonitis, aggravating an already severe condition and complicating the required surgery. The radiologist eventually read the plain film (but after the patient had been given oral contrast), and alerted the ED to the problem. The surgeons were then called, and the patient was taken to the operating room where a perforated ulcer with extensive peritonitis was repaired successfully. Post-operatively the patient suffered a severe stroke; the relationship of this to the preceding events is unclear.

Ironically, a meeting to investigate the cause of this patient's injury was held on the date of Case 1, prompting one of the participants to remark, "… the same thing is happening out there again today."

ANALYSIS

These cases represent episodes where the 'horizon of tractability' was exceeded by conditions beyond the range of previous operating experience. The resources and coping strategies that would normally provide resilience against variation and the unexpected became exhausted, compelling workers to invent new strategies 'on the fly' and to make sacrifice decisions, abandoning lower level goals in order to preserve higher ones and regain control of the situation (Cook & Nemeth, 2006).

Resilience State Space Model

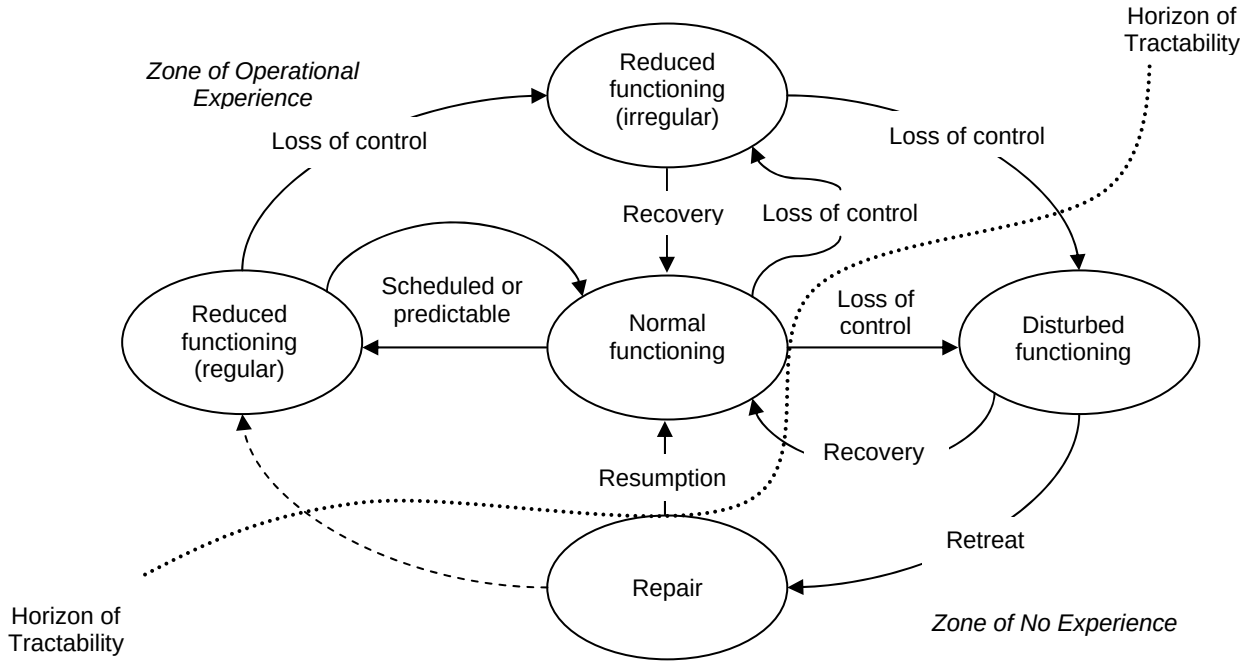


Figure 1. State-space diagram for service organizations. (Modified from Hollnagel & Sunström 2006, p 341, used with permission)

Adaptations in the Zone of Operational Experience

When the number of critical and serious patients needing assessment and intervention grew rapidly, (and seemingly without limit), the ED shifted to 'irregular reduced functioning'. This was marked by an attempt to continue with diagnostic and

therapeutic measures in all patients, using irregular spaces and informally supported sacrifices of some routine procedures. One interesting example here was the strategy of placing patients in chairs. It was never spoken explicitly, but widely recognized, that the ability to maintain postural tone (*ie*, to sit in a chair) was an indicator of a certain level of stability; thus management of patients in chairs could be sacrificed in order to attend to patients of higher criticality. In effect, this strategy identifies patients who might be physiologically more resilient, and "borrows" some of their resilience to provide additional capacity to support higher level goals and operations. Essentially this phase is characterized by attempts to develop compensatory buffers to help manage the disturbance.

A second adaptation involved sacrificing some lower level goals in order to be able to satisfy higher ones. For example, a national standard has been proposed that any chest pain patient should receive an electrocardiogram (ECG) within 15 minutes of arrival. In this ED, due to chronic decompensation, the mean time to ECG was typically around 35 minutes; in Case 1, the mean increased to 52 minutes (range 0 to 154 minutes), as workers concentrated on what they perceived as higher priorities.

A third adaptation was an anticipatory, 'feed-forward' strategy for ordering tests. This strategy assumes that the current disturbance will be transient, so the goal should be facilitating those functions that will be important on resumption of more nearly normal operations. Physicians used a strategy of anticipatory test ordering to try to 'save time in the future'; *ie*, instead of selecting tests in series, specifically tailored to a patient's condition (which would require a detailed assessment for which there was no time), physicians would order a broad battery of tests in parallel, assuming that by the time the results came back (typically in several hours), they would have completed that detailed assessment and would thus know which results were not relevant. This offers obvious advantages over waiting to place the order, since the results would then be even further delayed. This can be viewed also as a strategy for shifting some of the overload to other parts of the organization, and is a mechanism by which the disturbance spreads. In Case 2, this strategy led to placing the order for the CT scan without first reviewing the plain film and other, simpler tests.

What characterized all 3 adaptive strategies is that they had been tried before under similar circumstances, and thus were the 'usual responses' to 'normal, natural troubles.'

Crossing the 'Horizon of Tractability'

In both cases, the situation eventually worsened to 'disturbed functioning', where novel and highly irregular resources were employed. For example, a small office adjacent to the treatment area was used to perform ECGs on patients who were waiting in the aisles or in chairs, because it had a door that could be closed for privacy. Similarly, a small closet normally used for storage of respiratory and advanced airway equipment was used as a blood drawing area. In Case 2, the novel adaptation of triaging newly arrived patients to the hallway when stretcher spaces were exhausted is an example of using novel spaces to maintain some (reduced) level of functioning.

Ultimately, the ED was forced to retreat entirely from any semblance of routine operations for all but the most time-critical of patients. Essentially, this was a strategic decision to concentrate stop operations and regroup, and was manifested by a shift in operations from medical content to simple tracking – identifying patients, the (irregular) spaces to which they were assigned, and a vague categorization of problem type. In both cases, this was aided by creating a second status board within the ED's main status board. This second board was used for patients without assigned treatment areas who were waiting in chairs around the nursing stations, and listed only patient's name, location (this required some informal inventions, *eg*, 'Pyxis chair 2') and check boxes indicating that a physician had spoken to them, or that blood had been drawn. This is essentially the 'repair' state in Figure 1, and involves discontinuing operations in an attempt to regain control. In terms of goal states, it sacrifices most lower and intermediate level goals in order to preserve resources to restart the system once the disturbance had passed. (It is undoubtedly not accidental that this strategy is expressed in the rhetoric of defeat and resignation).

Once the repair had been successfully accomplished (in that workers now knew which patients they had responsibility for, where those patients were physically located, and what their basic problem type was), and the system stabilized (aided by the decrease in the numbers of incoming critical patients), then normal operations could be gradually resumed. This was done cautiously; it took some time to build up confidence that the current assessments were accurate and complete – the "continuing expectation of future surprise" (Rochlin, 1999) led to a conservative and gradual re-starting of routine operations.

The rapidity of the degradation in performance suggests that the ED possesses highly nonlinear characteristics. The flow of patients through the department on these days seems analogous to phase shifts in the state of matter; discontinuous transitions from laminar, to turbulent flow, to complete stagnation, similar to the condensation of water from a vapor, to a liquid, to ice.

DISCUSSION

These cases illustrate a complex pattern of performance degradations: acute decompensation, superimposed on chronic (Miller & Xiao, 2006). The ability of the staff to compensate during the period of chronic decompensation masked the drift toward the boundary of failure. This proximity to failure was finally revealed when buffers that were not easily further expanded were exceeded. Specifically, the lack of available physical space became the irreducible constraint in both cases that led the system ultimately to transition to the repair state.

Clinicians who self-select to work in EDs have a high tolerance for uncertainty, and take great pride in their ability to respond resiliently to uncertain and unpredictable demands. In terms of patient load, the demands in both these cases were not extraordinary; the total daily visits on these days were not above average, and the acute care unit had successfully managed mass casualty incidents – large numbers of critically ill or injured patients arriving simultaneously or in rapid succession – on numerous occasions in the past. Therefore, the sensation of "free fall" experienced on these two days was highly distressing to the health professionals involved. Rather than being able to "take things in one's stride", as they normally expect based on experience in managing the expected, normal and natural troubles, they were confronted instead with an acute sense of overwhelming failure and loss of control (Cook & Nemeth, 2006). Although they did not have the language of the resilience state space in which to express it, the distress that many senior, experienced workers felt over these incidents likely stems from this being their first, ever, transition into the repair state. Since by definition, an ED should never be in the repair state, such a transition challenges the very core of their professional identity. In addition, the impression that these episodes were related to hospital management issues, rather than external events (such as a hurricane or other disaster), added a sense of abandonment, which increased the affective impact on the workers.

CONCLUSION

Resilience in circumstances beyond operation experience is dynamic and adaptive, but finite in capacity. When the operational state leaves the zone of 'normal, natural troubles', workers shift on-the-fly to progressively more extreme and untested strategies in an attempt to compensate. If successful, some of these novel strategies may be adopted into the repertoire of usual responses, thus extending the capacity of the system. However, some novel strategies are associated with failure and are rapidly abandoned; the most novel strategy of all – retreat and repair – was successful but distasteful, as it challenged notions of professional competence and identity.

REFERENCES

Cook, R. I., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. *Quality & Safety in Health Care, 14*(2), 130-134.

Cook, R. I., & Nemeth, C. (2006). Taking things in one's stride: cognitive features of two resilient performances. In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 205 - 221). Aldershot, UK: Ashgate.

IOM Committee on the Future of Emergency Care in the US. (2006). *Hospital-Based Emergency Care: At the Breaking Point*. Washington, DC: National Academies Press.

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science, 42*(4), 237 - 270.

Miller, A., & Xiao, Y. (2006). Multi-level strategies to achieve resilience for an organisation operating at capacity: a case study at a trauma centre. *Cognition, Technology & Work*, 1-16.

Rochlin, G. I. (1999). Safe operation as a social construct. *Ergonomics, 42*(11), 1549-1560.

Vo**ß,** A., Procter, R., Slack, R., Hartswood, M., & Rouncefield, M. (2006). Understanding and supporting dependability as ordinary action. In K. Clarke, G. Hardstone, M. Rouncefield & I. Sommerville (Eds.), *Trust in Technology: A Socio-Technical Perspective* (pp. 195 - 216). Dordrecht, NL: Springer.

Wears, R. L., & Perry, S. J. (2006, 8 - 10 November 2006). *Free fall - a case study of resilience, its degradation, and recovery, in an emergency department.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Woods, D. D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL: CRC Press / Taylor & Francis Group.

Chapter X

Resilience in the Emergency Department

Robert L .Wears, Shawna J. Perry, Shilo Anders & David D. Woods

Introduction

Hospital emergency departments (EDs) are complex, dynamic settings where successful and effective work must occur in the face of high consequences of failure, and where practitioners are operating under conditions of time and resource constraints, physical stress (noise, fatigue), uncertainty, engaged in a multiplicity of tasks and resolving competition among goals, many of which are ill-defined, shifting, or ambiguous. ED work is made even more difficult because it is inherently limited to reacting to events – there is no possibility of seizing the initiative and controlling the pace of events (*eg*, as a military unit might do by going on the attack), or even of preparing for impending events in anything other than the most general way.

EDs are additionally interesting because, like many complex adaptive systems, they are historically emergent. EDs began to appear in hospitals in the 1950s, initially as simple loading zones where ambulances could deliver accident victims. Their evolution to their current state did not result from high level health or public policy planning (in fact, some segments of this community would prefer they went away), but rather was the cumulative result of decisions made by individual agents based on local knowledge and opportunity (Zink, 2006). Emergency care is now an established specialty field, with its own training programs and qualifications systems, and its own characteristic way of approaching the problems of illness and injury, most notably manifested in a shift of focus from an interest in understanding and predicting what will happen (typical of traditional medical and nursing activity), to a focus on identifying and forestalling what sorts of unfavourable events might happen.

For the most part, EDs perform their work in a remarkably resilient and adaptive way, such that events that might easily lead to a catastrophic outcome are often little more than small perturbations in the flow of high tempo events (Dismukes, Berman, & Loukopoulos,

2007). But, the resilient capacity of EDs is finite, and they are under increasing pressure as their adaptive resources are being consumed in responding to increasing complexity, growing demand, and shrinking resources. EDs are now in the paradoxical condition of having a fundamentally resilient nature, but becoming the primary locus of brittleness in the overall healthcare system. These new demands can lead to coordination breakdowns at boundary conditions. In this chapter, we analyze an active emergency department in terms of resilience concepts, in particular, to evaluate and illustrate descriptive models of resilience (*eg*, Hollnagel and Sundström's concept of resilience state space (Hollnagel & Sundström, 2006), Woods & Wreathall's analogy to physical materials under stress (Woods, Wreathall, & Anders, 2006), and Cook's description of the dynamics of resilient performance (Cook, 2006). The data are based on observations of an emergency department as it handles different loads and retrospective analyses of actual cases of situations that drove this system to its limit of adaptive capacity.

EDs seem to use four fundamental types of adaptive strategies in coping with the challenges of their work. A routine day is one in which the system is operating under usual conditions and described by practitioners as "run of the mill" where the system anticipates shifts beyond the routine and adapt apparently seamlessly. This would seem to fall into the normal functioning regions of the resilience state space (see Figure 3), or 'elastic region' of the stress-strain curve (see Figure 4).

In a second class of situations, a key person recognizes system degradation as load and demands are increasing, and initiates adaptive responses (*eg*, identifying and reorganizing additional resources, such as additional buffering capacity) to manage the challenges and maintain performance at near normal levels. Adaptations in these two settings are readily available solutions to the "expected, normal and natural troubles" with which workers have become familiar through experience, and word-of-mouth (Voß, Procter, Slack, Hartswood, & Rouncefield, 2006). For the most part, these adaptations are skilfully and unconsciously, almost invisibly performed, as expressed in the 'Law of Fluency' (Woods & Hollnagel, 2006). They are the usual solutions to the usual problems, and thus are contained with a 'horizon of tractability' (Voß et al., 2006).

In more extreme situations, the demands increase to the point that the required adaptations occur at the level of the whole department. In this case, the demands on the organization may cross the 'horizon of tractability' and ultimately challenge its ability to sustain operations and risk escalating to a breaking point, which has been described by practitioners as a 'free fall' (Wears & Perry, 2006). Practitioners have to recognize and anticipate the trend, and to reorganize activities and resources at the same time as they are struggling to handle patient load. The final class is qualitatively different from the three ordered classes mentioned, in that it involves planned for but rarely experienced events requiring a complete reorganization of work in the wake of a catastrophic event, such as a mass casualty event or natural disaster. For a variety of reasons, healthcare organizations are reluctant to shift to this 4th strategy in the absences of an unambiguous external trigger.

The chapter next provides a brief description of the ED setting, emphasizing characteristics generalizable across many EDs, followed by two illustrative case studies, analyzed in terms of the resilience concepts outlined above.

Setting

EDs are well-defined physical units in hospitals, but are ill-defined and fundamentally open systems in the functional sense. Because the physical span of control of ED workers is limited to reasonably small distances (say, less than 100 feet), very large EDs such as the one used as the data source for this chapter are typically subdivided into smaller units which are often functionally differentiated. For example, this ED is divided into five contiguous units, for trauma care, paediatric care, severe illness, and mild illness, with one unit reserved simply to hold admitted patients ('boarders') for whom there is no available bed in the hospital. The events described here took place in the five bed trauma unit and the 21 bed acute care unit of the ED; these units are physically adjacent and are staffed by mostly separate groups of nurses, but largely the same set of physicians who flow back and forth between units (see Figure 1).

ED workers consist of three professional groups – physicians, nurses, and technicians – who have a strong sense of identity and a distinct sense of a gradient in authority. (Other groups are often present in the ED but typically do not self-identify as ED practitioners, and often do

not work exclusively in the ED). These groups must coordinate their work, but act in highly independent manners, at a 'cooperative distance;' coordination among workers is largely implicit, mediated in part by external artefacts such as the status board, synchronous and asynchronous communication, and cross-monitoring.

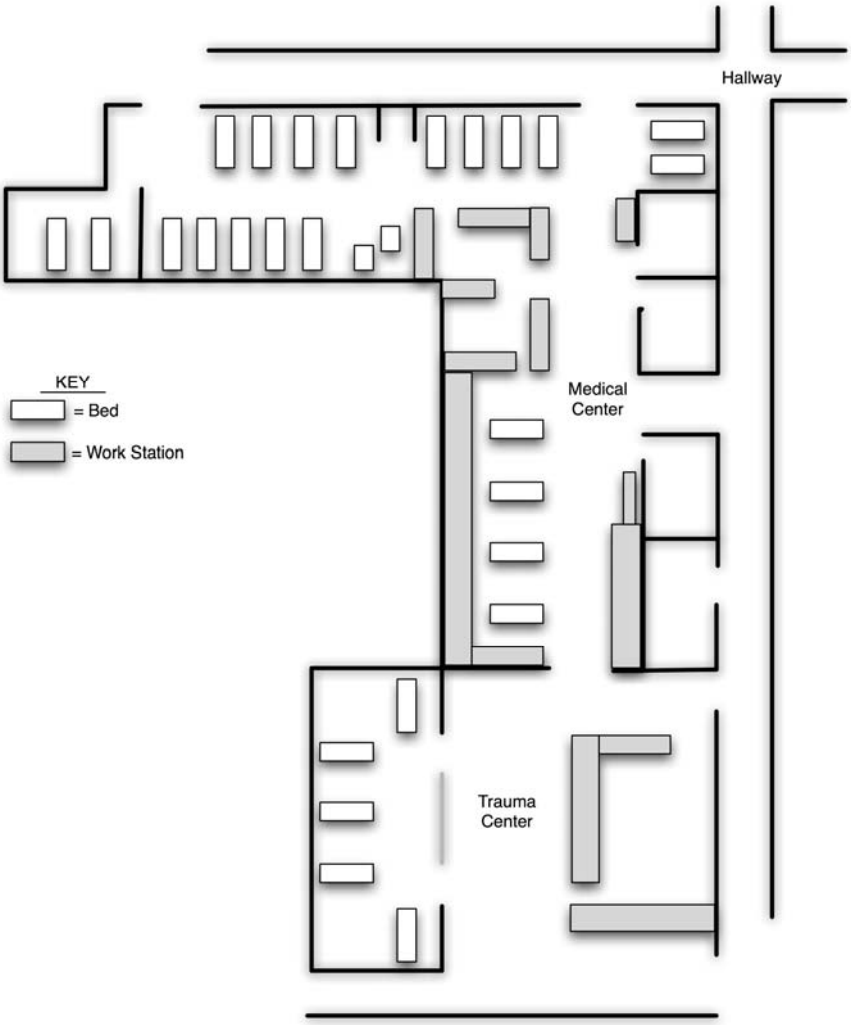


Figure 1. Schematic layout of the two units involved in these events.

Although it is seldom explicitly acknowledged, EDs provide important buffering and filtering functions for hospitals. The general reduction in US hospital capacity over the last 10 years, coupled with a gradual

increase in the number of ED patients due to population increase and aging, have led to overcrowding in many, if not most EDs(Richardson, Asplin, & Lowe, 2002; Schull, Vermeulen, Slaughter, Morrison, & Daly, 2004). This is typically manifested by the practice of holding admitted patients in the ED when there are no available in-patient beds in other, more suitable, parts of the hospital. This practice severely impacts ED operations (Schull et al., 2004; Sprivulis, Da Silva, Jacobs, Frazer, & Jelinek, 2006; Trzeciak & Rivers, 2003), but allows other parts of the hospital to continue functioning normally.

Case Studies

Case 1: Normal, natural troubles.

Before the escalating event of interest occurred in the emergency department, the night seemed to progressing in what could be described a 'run of the mill' fashion. The attending spent time shifting patients and deciding where to send the less critical patients in order to free up space in the units. Throughout the evening, a steady flow of patients was treated in both units under observation. The medical unit had only one critical bed occupied, while the trauma unit had received a number of patients earlier as well as from the shift before, hence it only had one open bed. The patients were all stable and were waiting to be transferred to other areas of the hospital. Of these patients, two were on ventilators, while the other two were conscious. This was the setting for the following case, which is described in a linear fashion with *intercalated comments* about the properties of resilience.

The trauma unit received a call about three incoming patients. To accommodate these patients, one current patient was transported to an in-patient bed, and another was moved to the hall.

Now the unit can only handle one more patient without reconfiguring. Therefore, they are too close to the margin if all three anticipated patients arrive, given the current capacity. They reconfigure by moving one patient upstairs and moving one patient to an area where ventilators cannot be used.

Patient #1 (first expected of three patients) arrived at the trauma center and was very combative due to head trauma; it took about 8 people to physically restrain him so he could be sedated to allow control of the situation and for diagnostic and therapeutic measures to proceed.

By using the relatively large resource of eight people now to sedate the patient, he will require less active monitoring later.

Two more patients arrive. The first is the second expected patient (#2) of three. The second (i.e., #3) is her child, who was not expected. The first is put in the open bed, while the child is taken to the paediatric unit of the ED. The paediatric fellow who took over care of the child had come to the trauma unit in response to a standard page given to all physicians when critical patients are arriving, but had not been aware that a paediatric patient was expected.

In order expedite the care of newly arrived critical patients, a set of attending and resident physicians are paged for any critical patients. When an unexpected child arrived, rather than helping in the trauma unit, the paediatric fellow changes plans, taking the child to the emergency paediatric unit herself, simultaneously preserving additional capacity in the trauma unit.

The unit is alerted that the third expected critical patient should arrive in less than five minutes. The attending asks the observer to get the chief resident from the acute care unit to help. The least critical of the remaining patients is wheeled into the hallway (next to the two patients already in the hall). The first patient is intubated and second patient is assessed.

The buffering capacity is increased by creating more beds before they are needed. In addition, this is a better buffer in that the space created allows the use of ventilators, which is not possible in the hallway.

Patient #4 arrives from an unrelated accident. The charge nurse asks the paramedic to page the nurse manager to get additional nursing staff. This patient is intubated at the same time as patient #1. The surgical attending arrives to decide which patient should be operated on first.

The charge nurse realizes that the trauma unit's resources (nursing staff) are running out. She unsuccessfully attempts to access resources from a larger resource pool (nursing for the entire hospital) by paging the nurse manager. This is a cross-scale interaction attempt to find additional resources in order to increase the distance between the current state of the system and the safety boundary. The surgical attending is opportunistically deciding which patients would benefit most from surgery, which will also frees up trauma unit resources.

The attending physician asks the radiology resident that is in the trauma unit to carefully examine all of the x rays and report any abnormal

findings to the trauma attending in order to minimize missing anomalies.

Attending physician realizes that in this state it is likely that an important alert might be missed, so she recruits other resources as a checking mechanism.

Patient #5 (husband of mother and child from car accident) arrives. All of the beds are taken and no more patients can be put into the hallway without blocking access. The attending physician asks the trauma charge nurse which patient requires fewer resources or is most stable and could be moved to a 'borrowed' bed in the adjacent acute care unit.

The trauma unit is reaching a boundary in that it has no more resources available within the unit itself, so in order to avoid collapse, the system shifts to utilization of resources in the acute care unit.

Patient #6 arrives with a knife wound. He is quickly examined and the charge nurse has the paramedics wait with the patient on the stretcher in the corner of the room until they have time to process him.

Personnel from outside the emergency department are pressed into service to monitor the patient in a holding pattern.

Patient #1 is taken to CT scan, and patient #5 is moved from the stretcher to a bed. The new critical patient #7 (hip fracture from car accident) arrives before a bed is made available, so ends up taking the space of the patient getting a CT scan. Patient #5 is prepared for a chest tube.

In all, 24 caregivers are in a small, noisy space, primarily caring for patients #5, #6, and #7. Although there are many patients, most resources are dedicated to a small number of prioritized patients. Bed and staff resources are flexibly recruited from other units, including the medical and paediatric unit. This recruitment signals an understanding that the situation is precarious in the sense that they are near the edge of what they can tolerate with current resources.

The medical charge nurse starts triage and intake of patient #8 (intoxicated patient who had driven into a telephone pole) in the hallway. Another nurse from the acute care unit assists the trauma nurses with patient care.

Facilitation occurs flexibly by sharing resources across the trauma and medical units.

Finally, 2 more patients (knife wound and bleeding from artery due to an accidental wound) walk into the trauma unit. A medical ED bed is recruited for their treatment by two resident physicians. Three

additional patients with minor wounds are stitched sequentially. Patient treatment continues without further incident for all other patients.

The system has returned to normal functioning.

Case 2: Beyond the 'Horizon of Tractability'

At the start of the evening shift (15:00), the ED was boarding 43 patients; 28 of these filled the unit reserved for boarders, the remaining 15 were split among the other two areas and the hallway separating the units. Seven were held in the hallway; all four of the acute care unit's critical care bays were filled with admitted patients on ventilators. As the shift change rounds began, the ED received a critically ill ambulance patient. Over the course of the next four hours, an additional five critically ill patients requiring ventilator support and other intensive measures arrived, in addition to multiple additional seriously but not critically ill patients (*eg*, chest pain suggestive of heart attack). All treatment spaces were filled; all temporary spaces to hold stretchers were filled; the unit ran out of stretchers and began 'storing' incoming patients in chairs near the nursing station. Congestion was severe, making it physically difficult to move around in the treatment area. This was particularly a problem when new critical patients arrived, since they needed to go to specific treatment spaces because of equipment requirements, and the patients occupying those spaces thus needed to be moved to other locations on very short notice. Figure 2 is a re-creation of the congestion experienced at the peak of this event.

The staff later described this situation as a feeling of *'free fall'*, a disorganized situation in which they did not know the numbers, types, or problems of the patients in their unit.

The crisis continued until approximately 22:00, by which time the staff felt they had finally gained control of the situation (in the sense of having a clear picture of which patients were present, where they were located, and at least a vague idea of the nature of their problem) and that the system had stabilized. No adverse events were associated with this episode, as far as is known.

Here, the 'horizon of tractability' (Voß et al., 2006)was exceeded by conditions beyond the range of previous operating experience. The resources and coping strategies that would normally provide resilience against variation and the unexpected became exhausted, compelling workers to invent new strategies 'on the fly' and to make sacrifice

decisions, abandoning lower level goals in order to preserve higher ones and regain control of the situation.

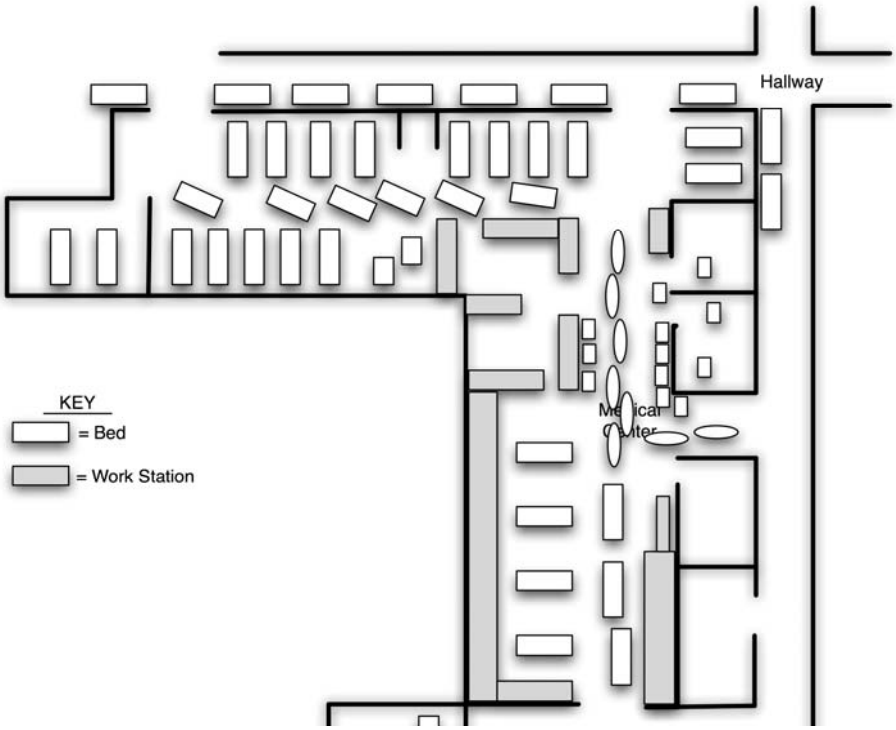


Figure 2. Re-construction of the congestion at the peak of the event described in Case 2. (Small ellipses represent ambulance patients on stretchers, waiting to be triaged; small rectangles represent patients in chairs).

Discussion

Resilience State Space Model

An abstract model describing the resilience state space – the set possible operating states and transitions among them – can be used to compactly describe these two cases (Hollnagel & Sundström, 2006). Figure 3 presents this model, as modified by us using in particular the concept of a 'horizon of tractability' (Voß et al., 2006), dividing the state space into two zones: one in which the usual solutions to the usual problems apply, and another that is either beyond the bounds of

operational experience or in which the degradation is so severe that repair attempts are very difficult or impossible.

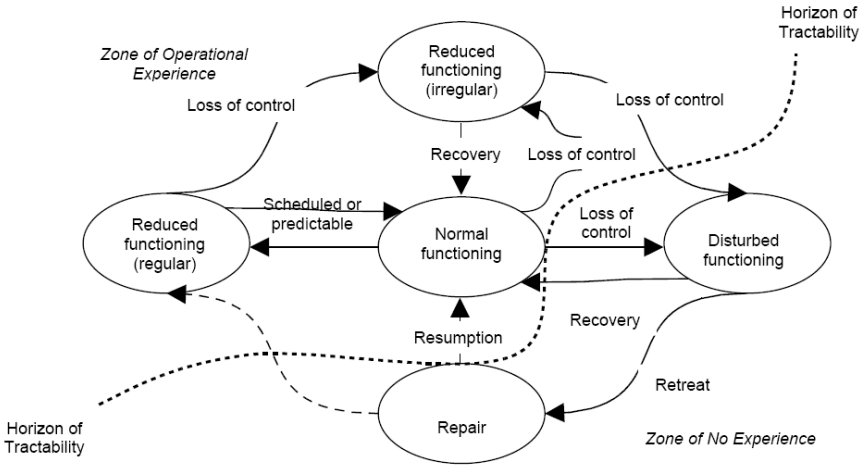


Figure 3. Resilience state space model (Hollnagel & Sundström, 2006), also illustrating the division into tractable and intractable zones (Voß et al., 2006).

Adaptations in the Zone of Operational Experience. Case 1 is characterized by state transitions that remain within the 'horizon of tractability'. The shift began in the normal state. (Here we use 'normal' to mean 'typical' or 'usual', as the institution had experienced a chronic degradation of operations over a period of years, as indicated by the presence of admitted patients in both units. This would have been viewed as highly irregular a few years prior, but had become the generally accepted operating condition. *Cf* 'normalization of deviance' (Vaughan, 1996).) A rapid increase in demand shifted the system into a variety of reduced functioning states, but rapid adaptations by actors in the field of practice were successful in keeping the situation manageable (*ie*, not crossing the 'horizon of tractability'), and regaining control and returning to the normal functioning state.

In Case 2, when the number of critical and serious patients needing assessment and intervention grew rapidly, (and seemingly without limit), the ED shifted to 'irregular reduced functioning'. This was marked by an attempt to continue with diagnostic and therapeutic measures in all patients, using irregular spaces and informally supported sacrifices of

some routine procedures. One interesting example here was the strategy of placing patients in chairs. It was never spoken explicitly, but widely recognized, that the ability to maintain postural tone (*ie*, to sit in a chair) was an indicator of a certain level of physiological stability; thus management of patients in chairs could be sacrificed in order to attend to patients of higher criticality. In effect, this strategy identifies patients who might be physiologically more resilient, and "borrows" some of their resilience to provide additional capacity to support higher level goals and operations. Essentially this phase is characterized by attempts to develop compensatory buffers to help manage the disturbance.

A second adaptation involved sacrificing some lower level goals in order to be able to satisfy higher ones. For example, a national standard has been proposed that any chest pain patient should receive an electrocardiogram (ECG) within 15 minutes of arrival. In this ED, due to chronic decompensation, the mean time to ECG was typically around 35 minutes; in Case 2, the mean increased to 52 minutes (range 0 to 154 minutes), as workers concentrated on what they perceived as higher priorities.

A third adaptation was an anticipatory, 'feed-forward' strategy for ordering tests. This strategy assumes that the current disturbance will be transient, so the goal should be facilitating those functions that will be important on resumption of more nearly normal operations. Physicians used a strategy of anticipatory test ordering to try to 'save time in the future'; ie, instead of selecting tests in series, specifically tailored to a patient's condition (which would require a detailed assessment for which there was no time), physicians would order a broad battery of tests in parallel, assuming that by the time the results came back (typically in several hours), they would have completed that detailed assessment and would thus know which results were not relevant. This offers obvious advantages over waiting to place the order, since the results would then be even further delayed. This can be viewed also as a strategy for shifting some of the overload to other parts of the organization, and is a mechanism by which the disturbance spreads.

What characterized all three adaptive strategies is that they had been tried before under similar circumstances, and thus were the 'usual responses' to 'normal, natural troubles.'

<u>Crossing the 'Horizon of Tractability'</u>. In Case 2, the situation eventually worsened to 'disturbed functioning', where novel and highly

irregular resources were employed. (This transition is analogous to a phase shift in the stress-strain curve in Figure 4, where an organization moves from the elastic region to the plastic region). For example, a small office adjacent to the treatment area was used to perform ECGs on patients who were waiting in the aisles or in chairs, because it had a door that could be closed for privacy. Similarly, a small closet normally used for storage of respiratory and advanced airway equipment was used as a blood drawing area. In another case not presented here, the novel adaptation of triaging newly arrived patients to the hallway when stretcher spaces were exhausted provides an additional example of using novel spaces to maintain some (reduced) level of functioning.

Ultimately, the ED was forced to retreat entirely from any semblance of routine operations for all but the most time-critical of patients. Essentially, this was a strategic decision to stop operations and regroup – a retreat into the 'repair' state in Figure 3. This transition was manifested by a shift in operations from medical content to simple tracking – identifying patients, the (irregular) spaces to which they were assigned, and a vague categorization of problem type. It essentially involves discontinuing operations in an attempt to regain control. In terms of goal states, it sacrifices almost all lower and intermediate level goals in order to preserve resources to restart the system once the disturbance had passed. (It is undoubtedly not accidental that this strategy is expressed in the rhetoric of defeat and resignation.)

Once the repair had been successfully accomplished (in that workers now knew which patients they had responsibility for, where those patients were physically located, and what their basic problem type was), and the system stabilized (aided by the decrease in the numbers of incoming critical patients), then normal operations could be gradually resumed. This was done cautiously; it took some time to build up confidence that the current assessments were accurate and complete – the "continuing expectation of future surprise" led to a conservative and gradual re-starting of routine operations.

Similarly, in case 1, the surgical residents completed several minor repairs in a bed in the acute care unit of the ED, even after the influx of patients had subsided in the trauma unit. As a result, the ED experienced the temporary loss of one critical patient bed; however, this conserved and increased the adaptive capacity of the trauma unit.

This slow return to normal, or gradual restarting, is roughly analogous to the physical phenomenon of *hysteresis*, where a system changes under some external influence, but as the influence is removed, the return to previous states is delayed; *ie*, it does not retrace in recovery the trajectory it traced in degradation.

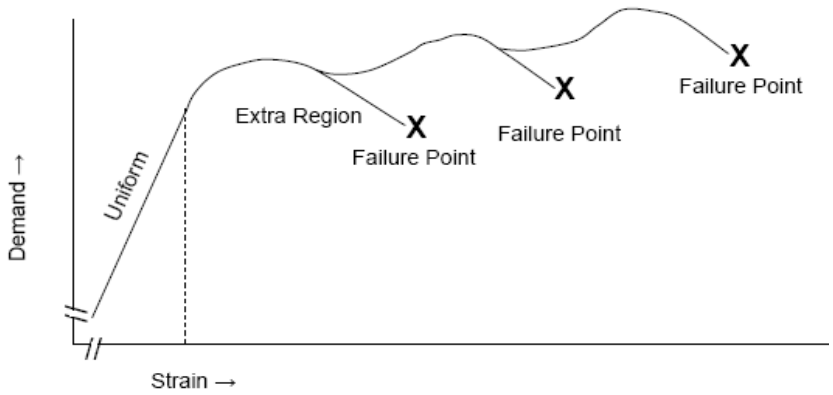


Figure 4. Stress-strain representation of performance. The region to the left of the vertical line represents elastic response; the region of plastic deformation is to the right of the line. (Modified from Woods, Wreathall & Anders, 2006).

Stress-Strain Curve Model

Figure 4 provides an alternative representation of resilience has been proposed, based an analogy to the stress-strain curves common in materials science (Woods *et al.*, 2006). In this analogy, the straight line marked 'uniform' on the left of the graphic represents the 'elastic' region of the system. Here, the normal, natural responses to normal, natural trouble work to allow the system to smoothly respond to demand. When routine adaptive capacity is exceeded, then the system enters the plastic region and 'deforms' – new adaptations and reconfigurations allow production to continue to meet demand, but at greater cost (in resources, effort, speed, and/or quality). Progressively increasing demand leads to new adaptations and reconfigurations (deformations) until ultimately adaptive capacity is exhausted and the system fails (the material fractures). These deformations / adaptations correspond roughly to the state transitions in Figure 3.

The stress-strain analogy is appealing because it suggests possible empirical application (Woods et al., 2006). If reliable and valid measures of demand and resource investment can be obtained, then several measures might be useful. The slope of the elastic region would represent the normal performance capacity of the system. The level of demand at which elasticity is lost would be its maximal normally tolerated demand. The average slope in the multiple deformation region reflects the adaptive capacity of the system. And of course, the level of demand at the point of failure might be estimable.

In addition, one might expect the quality of performance (as distinguished from the volume) to show a pattern that would be related to the first derivative (the slope) of the curve in Figure 4 (see Figure 5). In the elastic region, the quality of performance is uniform and constant (*ie*, the slope of the elastic, linear portion of the curve is invariant). When, for example, buffering capacity is exceeded and the system reconfigures on entry into the plastic region, quality begins to decline in a trade-off between quality and volume.

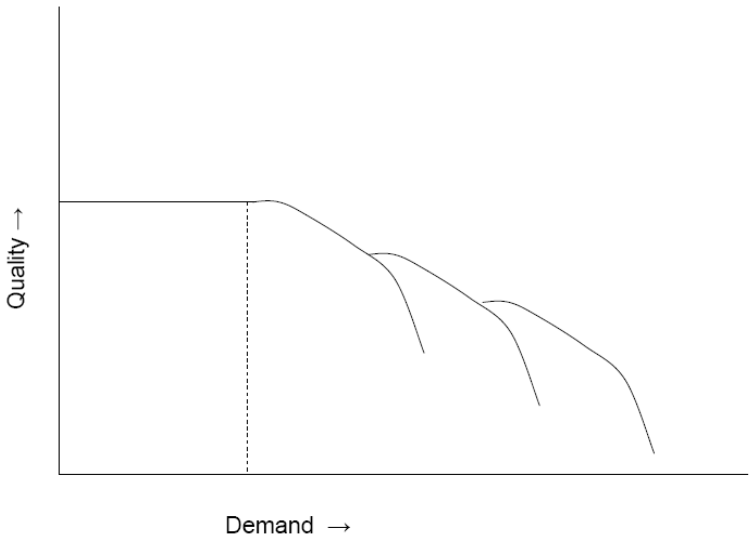


Figure 5. Effect of demand on performance quality. (Curves here represent the slope of the curve in Figure 4). In the elastic region (left of the dashed line), quality of performance is maintained in the face of

increasing demand, until a threshold is reached, corresponding to entry into the plastic region.

Some forthcoming work on the relationship of quality of performance in the ED to the level of demand suggests an interesting extension of this representation (see Figure 5) (Fee, Weber, Maak, & Bacchetti, 2007; Gray & Baraff, 2007; Pines & Hollander, 2007; Pines et al., 2007). Studies of performance in different clinical problems (time to antibiotic administration in pneumonia or serious bacterial infections, and time to analgesic administration in acute painful conditions) and in different institutions have shown an essentially linear relationship, without any evidence of a 'threshold' level, beyond which quality degrades; they suggest that some degradation in quality occurs at every observed increase in demand, even a starting points where the ED is not considered overloaded. This absence of the threshold effect postulated by Figure 5 suggests that EDs are currently almost always in the plastic region (or the irregular reduced function state).

Resilience dynamics

Neither of the two representations discussed so far directly includes a temporal dimension. Cook has noted that performances described as resilient may have different time dynamics, and has suggested a number of prototypical patterns (Cook, 2006).

In Figure 6, the light, upper boxes represent demand (D) and the dark, lower boxes represent the system's response (R), with time moving from left to right. Pattern 1 represents performance in the elastic (uniform) region in Figure 4, or the normal and/or regular-reduced states in Figure 3; here performance smoothly scales to meet demand. Pattern 2 represents the deformation / adaptation regions of Figure 4, or the irregular-reduced or disturbed function states in Figure 3. Here, the adaptations have allowed a greater response to higher demand, so demand eventually increases (see the 'Law of Stretched System' (Hirschhorn, 1997)). Case 1 showed both patterns 1 and 2 in its evolution.

Pattern 3, which was illustrated by Case 2, shows that the capacity to adapt to increased demand is not infinite, and once exceeded, the system loses its ability to respond to demand: a new insult leads to failure, followed by halting recovery. What is important to note here is that the loss of 'last reserves' is inapparent (and indeed, the pattern may

have been interpreted as efficient, or even heroic, response) until the final shock and total collapse occurs. Patterns 2 and 3 also show the phenomenon of hysteresis, discussed above.

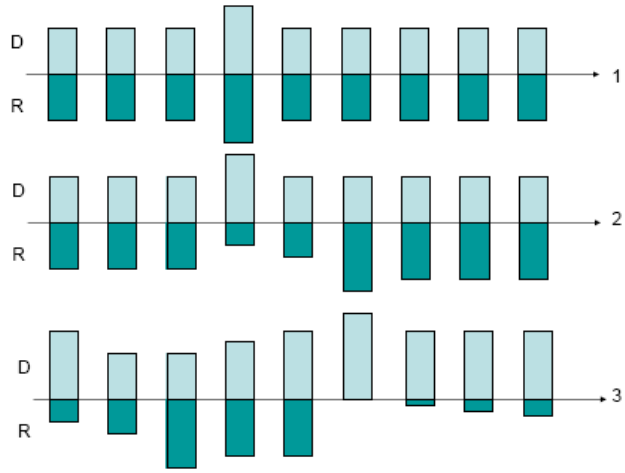


Figure 6. Three temporal patterns of resilience, illustrating elastic performance (1), deformation / adaptation (2), and failure (3). Modified from Cook, 2006.

Conclusion

These cases illustrate a complex pattern of performance degradations: acute decompensation, superimposed on chronic erosion of capacity. The ability of the staff to compensate during the period of chronic decompensation masked the drift toward the boundary of failure. This proximity to failure was finally revealed when buffers that could not easily be further expanded were exceeded. Specifically, the lack of available physical space became the irreducible constraint in both cases that led the system ultimately to transition to the repair state.

Clinicians who self-select to work in EDs have a high tolerance for uncertainty, and take great pride in their ability to respond resiliently to uncertain and unpredictable demands. In terms of patient load, the demands in both these cases were not extraordinary; the total daily visits on these days were not above average, and the acute care unit had successfully managed mass casualty incidents – large numbers of critically ill or injured patients arriving simultaneously or in rapid succession – on numerous occasions in the past. Therefore, the sensation of "free fall" experienced in Case 2 was highly distressing to

the health professionals involved. Rather than being able to "take things in one's stride", as they normally expect based on experience in managing the expected, normal and natural troubles, they were confronted instead with an acute sense of overwhelming failure and loss of control. Although they did not have the language of the resilience state space in which to express it, the distress that many senior, experienced workers felt over these incidents likely stems from this being their first, ever, transition into the repair state. Since by definition, an ED should never be in the repair state, such a transition challenges the very core of their professional identity. In addition, the impression that these episodes were related to hospital management issues, rather than external events (such as a hurricane or other disaster), added a sense of abandonment, which increased the affective impact on the workers.

Resilience in circumstances beyond operational experience is dynamic and adaptive, but finite in capacity. When the operational state leaves the zone of 'normal, natural troubles', workers shift on-the-fly to progressively more extreme and untested strategies in an attempt to compensate; they reposition themselves to a different curve in the 'plastic' region rather than reach a failure point. If successful, some of these novel strategies may be adopted into the repertoire of usual responses, thus extending the capacity of the system. However, some novel strategies are associated with failure and are rapidly abandoned. In Case 2, the most novel strategy of all – retreat and repair – was successful but distasteful, as it challenged notions of professional competence and identity.

References

Cook, R. I. (2006). *Resilience dynamics.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Dismukes, R. K., Berman, B. A., & Loukopoulos, L. D. (2007). *The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents*. Aldershot, UK: Ashgate.

Fee, C., Weber, E. J., Maak, C. A., & Bacchetti, P. (2007). Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia. *Annals of Emergency Medicine, x*(x), xx (in press).

Gray, Z. A., & Baraff, L. J. (2007). The effect of emergency department crowding on time to parenteral antibiotics in admitted patients with serious bacterial infections. *Annals of Emergency Medicine, x*(x), xx (in press).

Hirschhorn, L. (1997). Law of Stretched Systems, quoted in Woods & Cook. Retrieved 2 February 2005, from http://www.ctlab.org/properties/pdf%20files/Characteristics %20of%20Patient%20Safety.PDF

Hollnagel, E., & Sundström, G. (2006). States of resilience. In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 339 - 346). Aldershot, UK: Ashgate.

Pines, J. M., & Hollander, J. E. (2007). Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine, x*(x), xx (in press).

Pines, J. M., Localio, R., Hollander, J. E., Baxt, W. G., Lee, H., Phillips, C., et al. (2007). The impact of ED crowding measures on time to antibiotics for patients with community-acquired pneumonia. *Annals of Emergency Medicine, x*(x), xx (in press).

Richardson, L. D., Asplin, B. R., & Lowe, R. A. (2002). Emergency department crowding as a health policy issue: past development, future directions. *Ann Emerg Med, 40*(4), 388-393.

Schull, M. J., Vermeulen, M., Slaughter, G., Morrison, L., & Daly, P. (2004). Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Ann Emerg Med, 44*(6), 577-585.

Sprivulis, P. C., Da Silva, J. A., Jacobs, I. G., Frazer, A. R., & Jelinek, G. A. (2006). The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Med J Aust, 184*(5), 208-212.

Trzeciak, S., & Rivers, E. P. (2003). Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. *Emerg Med J, 20*(5), 402-405.

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*. Chicago, IL: University of Chicago Press.

Vo**ß,** A., Procter, R., Slack, R., Hartswood, M., & Rouncefield, M. (2006). Understanding and supporting dependability as ordinary action. In K. Clarke, G. Hardstone, M. Rouncefield & I.

Sommerville (Eds.), *Trust in Technology: A Socio-Technical Perspective* (pp. 195 - 216). Dordrecht, NL: Springer.

Wears, R. L., & Perry, S. J. (2006). *Free fall - a case study of resilience, its degradation, and recovery, in an emergency department.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Woods, D. D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering.* Boca Raton, FL: CRC Press / Taylor & Francis Group.

Woods, D. D., Wreathall, J., & Anders, S. (2006). *Stress-strain plots as a model of an organization's resilience.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Zink, B. J. (2006). *Anyone, Anything, Anytime: A History of Emergency Medicine.* Amsterdam, NL: Elsevier.

A Systems Dynamics Representation of Resilience

Robert L Wears[1,2] and Shawna J Perry[3]
[1]Clinical Safety Research Unit, Imperial College London, UK
wears@ufl.edu; r.wears@imperial.ac.uk
[2]University of Florida, Jacksonville, FL, USA
[3]Virginia Commonwealth University, Richmond, VA, USA
sperry4@mcvh-vcu.edu

Abstract. Resilience can be thought of as a property of a system that permits it to survive and achieve its goals in the face of expected threats and challenges to its operations. Systems dynamics modelling is a technique useful in exploring the behaviour of complex systems, especially the nonlinear interactions and feedback delays are present. In this paper, we use systems dynamics modelling to explore the nature of resilience, using small, highly abstract modules built for incorporation into a larger model of the crowding problem in emergency departments.

1 INTRODUCTION

Emergency departments (EDs) are dynamic, open, high risk, continuously operating systems that demonstrate considerable resilient capacity (Wears & Perry, 2006; Wears, Perry, & McFauls, 2007), but occasionally perform in less resilient, more brittle ways (Anders, Woods, Wears, Perry, & Patterson, 2006). Systems dynamics is a family of techniques for representing and exploring the behaviour of complex systems and their response to change over time (Sterman, 2000a). The objective of this paper is to use systems dynamics modelling of the problem of ED overcrowding to explore the nature of the transitions from resilience to brittleness and back again.

Most US EDs have experienced severe and increasing over-crowding problems over the past decade (Derlet, Richards, & Kravitz, 2001; Goldberg, 2000; Richardson, Asplin, & Lowe, 2002). This is thought to be due primarily to a decrease in the total number of inpatient beds *via* hospital closures, mergers and acquisitions (although there are many other causal influences), leading to the 'boarding' of large numbers of admitted patients in the ED. EDs have adapted to this problem in a variety of ways, such as dedicating entire units to inpatients, adapting previously unused space such as hallways to use as treatment spaces, and dynamically changing the manner in which work is performed (Wears & Perry, 2006). These adaptations create a series of reverberations throughout the organisation that eventually feed back to affect the ED, although subject to various time delays. As the over-crowding problem has increased in severity, this adaptive capacity has become increasingly strained, and a highly respected study of the problem has concluded that EDs as a whole are near a point of complete breakdown (Committee on the Future of Emergency Care in the US, 2006).

Because of time delays in feedback and complex interactions with the rest of the hospital, the effect of current or proposed future strategies to maintain safe ED operations is difficult to determine; in fact, some of the proposed solutions may even be making the problem worse in the long run. The problem has largely been viewed as intractable, and has resisted many attempts at solution or mitigation.

This paper stems from a larger project to model the overcrowding problem and potential approaches to it in an attempt to provide policy guidance to organizations and managers. In this paper, we examine small, highly abstract modules that will be linked together as components of the larger ED model. The objective here is to characterize the sorts of model behaviours that might represent resilience or its converse, brittleness. Identifying these behaviours in very simple abstract models will be an important aid to assessing them in the more complex, 'full ED' that is currently under construction.

2 METHODS

In this section we describe the work system under consideration, then discuss the philosphy guiding model development, and finally describe the components of the model used for this analysis.

2.1 Work System

An urban 653 bed US teaching hospital that is part of an 8 hospital network served as the data source. The ED has roughly 100,000 visits per year, and is a Level 1 trauma center. It is divided into 5 major treatment areas totaling 79 beds; 2 treatment areas are dedicated to severe trauma patients and to pediatric cases. One of the non-dedicated treatment areas (comprising 22 beds) is reserved for 'boarders.' Two large hallways are routinely used as additional treatment space.

EDs are staffed by three distinct groups – physicians, nurses, and technicians – who have a strong sense of professional identity and a distinct sense of a gradient in authority. (Other groups also work in the ED but typically do not self-identify as ED practitioners, do not work there exclusively). These groups must coordinate their work, but act in highly independent manners, at a 'cooperative distance;' coordination among workers is largely implicit, mediated in part by external artefacts such as the status board (Wears, Perry, Wilson, Galliers, & Fone, 2007), synchronous and asynchronous communication, and cross-monitoring

2.2 Guiding Principles for Model Development

Many different modeling disciplines for this problem are available, and even within a single discipline there are a large number of choices to be made.

Balancing Scope and Detail. In evaluating this tradeoff, we decided to favor broad model boundaries over more extensive detail. A broader scope helps to avoid the problem of tacitly mistaking endogenous factors for external causes. We therefore bounded the model at the organizational level (the hospital), rather than at the departmental, or departmental unit level, because we wanted to explore the possible feedback relationships between the ED and the hospital.

With a broad scope, attempting to model fine-grained detail would become unmanageable, and in addition would limit the generalizability of the results.

Because two important goals of the project are to illuminate some aspects of resilience in many, not just this, ED; and even more broadly, to tease out some aspects of resilience in complex work systems in general, we purposely abstracted much of the underlying detail into simpler and hopefully more general model structures. For example, patients differ greatly in the amount of resources and effort they require of the ED, but we treat them as uniform; ED work is pulsatile, responding to at least 4 temporal cycles (daily, weekly, and seasonal cycles of visits and acuity, and weekly cycles of hospital bed availability), but we focus here on the averages. All of these assumptions will require empirical inspection and / or sensitivity analysis.

The temporal scope of the model is more limited than is typical for most systems dynamics models. We limit the temporal scope to the span of control typically wielded by ED operations and hospital middle managers, *ie*, to dealing with problems of flow and crowding in a span of days to weeks, not months to years; this essentially limits responses to reallocations of existing resources and priorities. Thus, increasing capacity by building new space, or hiring additional staff, are bounded out of this specific model because the time course for these actions is too long. Strategies at this level will be explored in related models but not dealt with further here.

Model Choice. The foregoing considerations led us to a generic modeling type, the compartment aggregation model, in which various states of the process are considered separate but communicating compartments; flows and levels are modeled, but not individual agents or entities. Two particular types of compartment aggregation models, aging chain models and supply chain models (Sterman, 2000b) are both well understood and particularly suitable for this problem. In an aging chain, material (or information) flows through a series of compartments in which it typically is delayed; the output of one compartment is the input into the next. Thus an aging chain model lends itself nicely to the logical progress of patients through various stages of the ED and on into the hospital. However, patients can also flow backward, or be recycled through the chain, so supply chain models, which allow for 're-work', are also attractive. One salient difference between the ED and supply chain models is that services cannot be provided in advance or excess and banked in inventories in the same way that 'widgets' can. Thus the ED model will be a hybrid of the classic aging chain and supply chain models (Orcun, Uzsoy, & Kempf, 2006).

Data. The model development process is continuing, and involves direct observations of ED operations, with special attention to 'limit cases' – situations of extreme congestion or overload. Quantitative data on numbers of patients, triage acuity, times of presentation and disposition are obtained from the hospitals computerized information system. Information on system performance was gathered from interviews with nursing, physician, ancillary and management staff in various departments of the organization (ie, not just the ED).

2.3 The Model and the Abstract Modules

A example model for studying the overcrowding problem is shown in Figure 1. It is a hypbrid of classical aging and supply chain models. ED patients present for care, undergo some processing (with delays), and eventually are admitted to the hospital or discharged. If admitted, they may be physically boarded in the ED or transported to an inpatient ward, depending on available inpatient beds. Poor discharge

decisions may lead to 'rework'; patients returning to a previous stage (by presenting again to the ED).

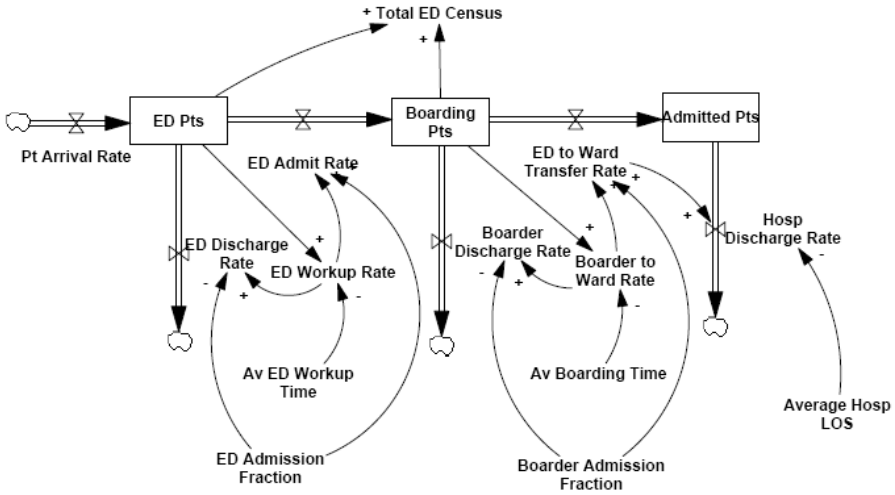


Fig. 1. Schematic model of patient flow through the ED and processing delays

Each of the 'processing units' shown in Figure 1 has a substructure (shown in Figure 2) and it is the general substructure and behaviour of those units that is the focus of this paper. These modules are highly simplified, capacitated, input-output units. Patients arrive at some rate, are processed, and leave at a rate that depends not only on the number of patients to be processed, but also on the relative proportions of patients to available resources, and on adaptations to workload by actors in the system.

2.4 Conjectures

If the model is to be useful, we postulate (hypothesize) that is should show several behaviours that seem characteristic of resilience.

Conjecture 1: Non-Adaptive Systems Show Brittleness. The simple modules developed as components for the larger system model should show some sensitivity to a set of inputs that produces sudden and dramatic changes in state.

Conjecture 2: Adding Adaptive Components Mitigate Brittleness. The addition of a capacity that adjusts to exogenous shocks should mitigate this brittleness.

Conjecture 3: Repeated Shocks Plus Adaptive Components Lead to 'Anticipatory' Compensation. If exogenous shocks are recurrent, and if the memory of adaptations is sufficiently strong, systems will migrate towards the adapted state and can respond more quickly to exogenous shocks.

Conjecture 4: Brittleness transmits, but resilience contains, exogenous shocks. When systems are arranged in chains, transmission of exogenous shocks along chains suggests brittleness, while their isolation in a small number of modules (ideally one) suggests resilience.

It is important to note here that overall performance of the ED model is likely to be emergent; *ie*, that resilient (or brittle) behaviour of the model as a whole need not necessarily be found in any of its components, and conversely that brittleness (or

resilience) at the component level may not necessarily imply that it exists for the entire model. The purpose here is to gain a better understanding of how those behaviors arise and how they might appear in the outputs by using the simplest possible modules.

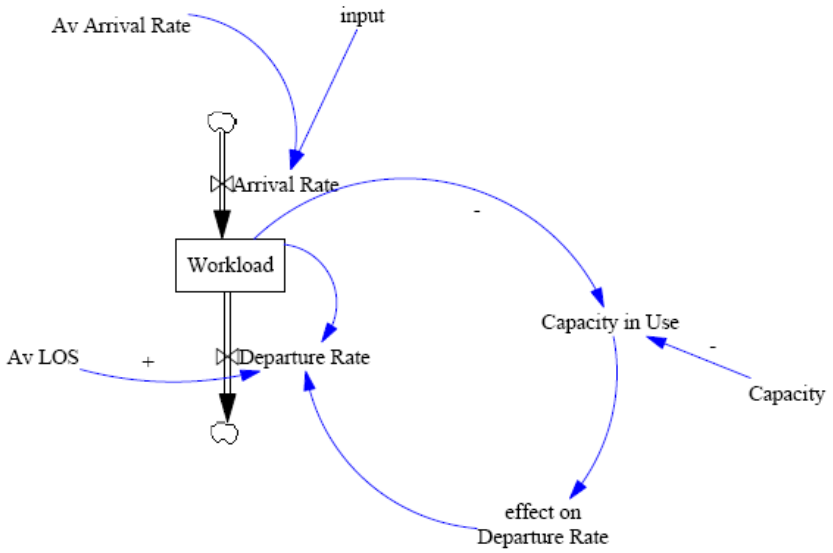


Fig. 2. Simple input-output component of ED model. This is a capacitated processing model; patients present at some rate (which may be modified by external shocks), undergo processing, and depart at a rate which is dependent on the number of patients and the capacity of the unit

3 RESULTS

The highly abstract modules shown in Figure 2 were simulated using historical data for on of the units of the ED, to evaluate the first three conjectures. Using historical data, the module exhibits steady-state performance (not shown) in that the number of patients in the unit stabilizes around 32, and inputs and outputs are balanced.

We first evaluated conjecture 1, that the system could show brittleness under certain conditions when adaptation was limited. Figure 3 shows this behaviour. When shocked by a single input pulse, system response increases to compensate, but eventually the system becomes overloaded and crashes catastrophically, even though the initial pulse was limited in time and input had returned to normal. It is interesting to note that this sudden degradation occurred at some time removed from the initial insult.

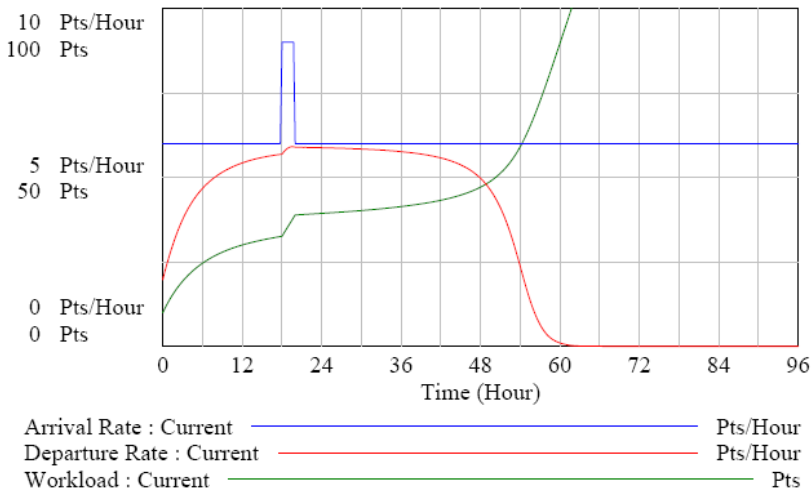


Fig. 3. Brittle response of the simple module to a single pulse load. Performance initially compensates, but eventually deteriorates catastrophically, some time after the initial shock

The addition of adaptive capacity to the model mitigates this response (Nathanael & Marmaras, 2008). Adaptive capacity is not further specified here – in a real world model it might take the form of work-arounds or short cuts that are employed when workers recognize overload and try to compensate for it. Figure 4 illustrates an adaptive, resilient response to a single pulse shock.

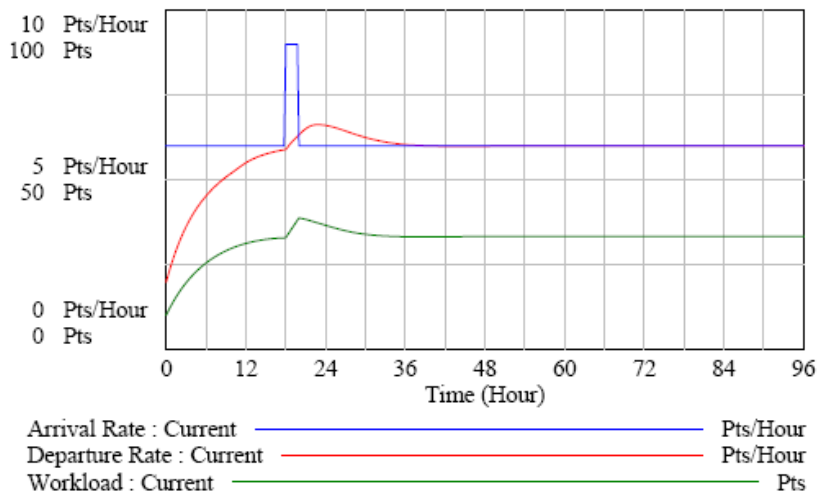


Fig. 4. Resilient response of the simple module to a single pulse load. Performance rises to compensates and gradually returns to the steady state.

With respect to Conjecture 3, the ability of the system to permanently adapt to repeated shocks by varying the rate of decay of adaptations. Figure 5 illustrates this effect for two rates of decay. After 3 pulse challenges, the two systems start to diverge, and the system with rapid decay (red line) undergoes a phase transition into catastrophic collapse, while the system with a longer 'memory' is able to shift to a new steady state at a higher load.

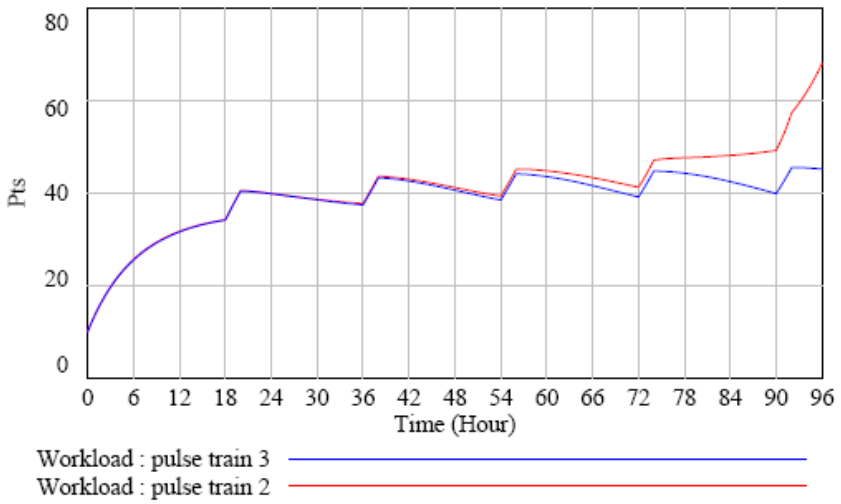


Fig. 5. Comparison of rapid (red) and slow (blue) decay of adaptations when challenged with repeated pulse shocks.

Finally, we evaluated Conjecture 4 by coupling the abstract ED model to a highly simplified hospital model (since outputs from the ED are one of the inputs to the hospital) as shown in Figure 6.

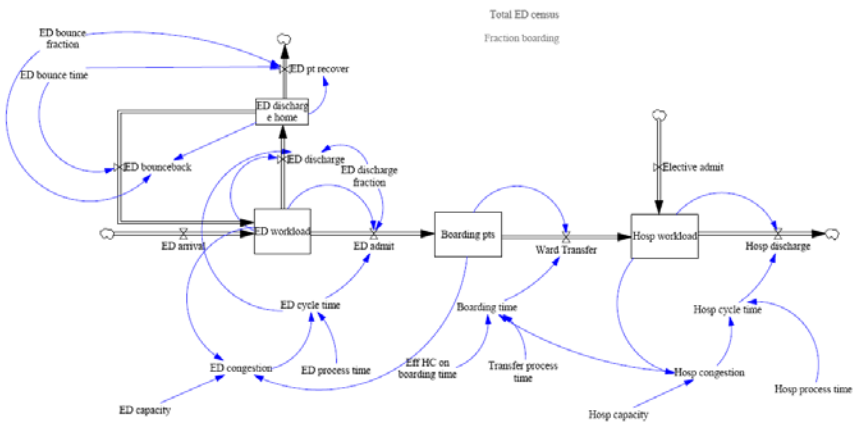


Fig. 6. ED model (left) coupled to hospital model (right), with feedbacks between each.

We then assessed the responses of these combined systems to a single external shock under varying values of ED capacity. Figure 7 demonstrates the disparate effect of changes in ED capacity. Increased capacity in the ED dramatically improves its ability to weather an external shock, and decreases in capacity dramatically degrade it. Conversely, improvements in ED capacity actually *worsen* hospital workload, while decreases in ED capacity have little to no effect on the hospital.

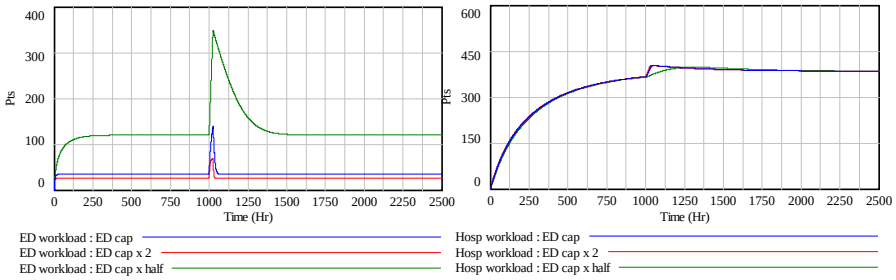


Fig. 7. Changes in ED (left) and hospital (right) workload occasioned by a single external shock, at varying levels of ED capacity.

4 DISCUSSION

"All models are wrong, but some models are useful" (Box, 1976). The models presented here are limited and highly abstract, but illustrate the possibility of obtaining more complex behaviours, such as resilience and brittleness from simple components. However, the models currently omit many important variables, such as fatigue, burnout, or the effect on quality (particularly the effect of rework as a result of diminished quality). Once refined and linked into a larger representation of the ED within the hospital, they may provide insight into a public policy problem that has so far resisted solution.

REFERENCES

Anders, S., Woods, D. D., Wears, R. L., Perry, S. J., & Patterson, E. S. (2006, 8 - 10 November 2006). *Limits on adaptation: modeling resilience and brittleness in hospital emergency departments.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791 - 799.

Committee on the Future of Emergency Care in the US. (2006). *Hospital-Based Emergency Care: At the Breaking Point*. Washington, DC: National Academies Press.

Derlet, R., Richards, J., & Kravitz, R. (2001). Frequent overcrowding in U.S. emergency departments. *Acad Emerg Med, 8*(2), 151-155.

Goldberg, C. (2000, December 17, 2000). Emergency crews worry as hospitals say, 'No vacancy'. *New York Times,* pp. Section 1, pg 27.

Nathanael, D., & Marmaras, N. (2008). Work practices and prescription: a key issue for organizational resilience. In E. Hollnagel, C. P. Nemeth & S. Dekker (Eds.), *Remaining Sensitive to the Possibility of Failure* (pp. 101 - 118). Aldershot, UK: Ashgate.

Orcun, S., Uzsoy, R., & Kempf, K. (2006). *Using System Dynamics Simulations to Compare Capacity Models for Production Planning.* Paper presented at the 2006 Winter Simulation Conference

Richardson, L. D., Asplin, B. R., & Lowe, R. A. (2002). Emergency department crowding as a health policy issue: past development, future directions. *Ann Emerg Med, 40*(4), 388-393.

Sterman, J. D. (2000a). *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin McGraw-Hill.

Sterman, J. D. (2000b). Co-flows and aging chains. In *Business Dynamics: Systems Thinking and Modeling for a Complex World* (pp. 469 - 512). Boston: Irwin McGraw-Hill.

Wears, R. L., & Perry, S. J. (2006). *Free fall - a case study of resilience, its degradation, and recovery, in an emergency department.* Paper presented at the 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France, http://www.resilience-engineering.org/REPapers/Wears_et_al.pdf

Wears, R. L., Perry, S. J., & McFauls, A. (2007). *Dynamic changes in reliability and resilience in the emergency department*. Paper presented at the 51st Human Factors and Ergonomics Society.

Wears, R. L., Perry, S. J., Wilson, S., Galliers, J., & Fone, J. (2007). Emergency department status boards: user-evolved artefacts for inter- and intra-group coordination. *Cognition, Technology & Work, 9*(3), 163-170.

Changing Horses in Midstream: Sudden Changes in Plan in Dynamic Decision-making Problems

Silvia Gilardi
Universitá di Milano
silvia.gilardi@unimi.it

Chiara Guglielmetti
Universitá di Milano
chiara.guglielmetti@unimi.it

Shawna J Perry
Virginia Commonwealth University
sperry4@mcvh-vcu.edu

Gabriella Pravettoni
Universitá di Milano
gabriella.pravettoni@unimi.it

Robert L Wears
University of Florida; Imperial College London
wears@ufl.edu; r.wears@imperial.ac.uk

ABSTRACT

Motivation – Why and how do expert unexpectedly change from their original plan in dynamic, uncertain settings? **Research approach** – Critical event interviews of practitioners self-reporting cases of sudden plan changes. **Findings** – Sudden plan changes developed in much the same manner as an expert's initial plan in recognition-primed decision-making. **Limitations** – Cases were limited to the healthcare domain; self-reporting distorts some aspects of the events; only cases where the change in plan worked were volunteered. **Originality/Value** – Many studies have examined persistence in an erroneous plan; fewer have examined sudden switches from bad to good plans. **Take away message** – Sudden plan changes arise in ways similar to experts' initial plan formulations: appearing as if unbidden, often not preceded by growing awareness of the need for reassessment, and once present seem obviously correct.

Keywords

Health care, dynamic decision problems, changes in plan, cognition and decision-making.

INTRODUCTION

Healthcare work typically involves dynamic decision-making problems (Brehmer, 1987), in which problems must be recognized and characterized, often as actions are being taken; these actions then affect the state of the world, thereby changing the existing problem, creating new problems, or re-ordering priorities. Although in their formal discussions of decision-making, health professionals seldom articulate this evolving, sense-making, dynamic aspect of clinical work, characterizing problems as fixed entities that can be resolved by applying the proper set of preplanned procedures, their actual practices do reflect it (although sometimes in an apologetic, *sub rosa* manner). One manifestation of this adaptiveness can be seen when clinicians rapidly change plans for the evaluation and management of patients. We report 4 cases in clinical decision-making involving a significant and sudden change of plans. The cases were elicited from practitioners in acute care settings using the critical incident method (Crandall, Klein, & Hoffman, 2006). The cases are presented without giving their final resolutions, to minimize the effects of outcome knowledge.

CASE SYNOPSES

Case 1

An extremely agitated and violent young man was brought into the emergency department (ED), shackled face down on a stretcher after having attacked a police officer following a minor traffic accident due to erratic driving. The police has used a Taser multiple times to try to control him. No other history was available. The ED staff began to follow their rapid sedation protocol, but after only a short time, decided to abandon the established protocol and instead use another drug, propofol, typically used for anesthesia; propofol use is strictly controlled because it can rapidly stop breathing and severely low blood pressure; its use in this setting violated organizational protocols set up when .

Case 2

An elderly man with chronic heart disease was admitted to the hospital for suspected pneumonia that failed to respond to outpatient antibiotics. Despite intensification of his antibiotic and cardiac therapies, he became progressively worse, and was transferred to the intensive care unit (ICU) and placed on a ventilator, but continued to deteriorate. Microbiological studies over the course of hospitalization did not identify an agent, and biopsies were compatible with, but not diagnostic of, pulmonary toxicity due to amiodarone, one of his cardiac medications. Amiodarone was stopped and a drug to treat the toxicity, prednisone was started; but prednisone is generally contraindicated in serious infections because it suppresses the immune system.

© The Authors 2009. Published by the British Computer Society

Case 3

A child was brought to the pediatric ED after having fallen onto a concrete bench. Because of the severity of pain, the physician was concerned about serious injury, but the initial ultrasound scan was negative. The ED nurse felt the child was only frightened and tried to get him discharged. Initial blood tests showed abnormalities, so the plan was to admit him for observation. The nurse began to feel uneasy, and asked for the physicians to reassess the child. They agreed to call a surgical consult, but took no other action. The child seemed to worsen so the nurse moved with him to the surgical ED, and then to CT scan, despite the fact that she was supposed to return directly to triage.

Case 4

A middle-aged man was seen in the trauma center after a motor vehicle crash, complaining of chest and abdominal pain. His trauma evaluation was negative for serious injury requiring surgery, and the trauma team's plan was to admit him to the hospital for observation. One physician noted the patient "didn't look like" a trauma case and ordered an electrocardiogram; it showed the patient was also having a heart attack.

COMPARISON AND CONTRAST

All 4 cases began as problems that were considered readily apparent, relatively straightforward, and for which well-understood and frequently practiced procedures existed. In Case 1, the clinicians' understanding of the problem did not change, but the plan dramatically changed to one that in fact involved a deliberate violation of existing procedures; thus it sheds light on the projective aspects of decision-making (since it entailed an expectation of failure if standard procedures were continued), and on the area of normal or necessary violations. The decision was not triggered by a a violation of expectations, but rather a mental simulation of projected possible courses (all of which were bad). Case 2 did not involve a violation of formal policies, but is similar since the change in plan was contradictory to the original plan, because immunosuppressing drugs are relatively contraindicated in suspected infections.

In Case 3, the change in understanding of the problem was not initially shared across the entire workgroup; thus it exposes the negotiated nature of problem recognition and characterization. The nurses' understanding here completely reversed itself, and it led her to violate expected procedures. Case 4 was characterized by a similar sudden change in the understanding of the patient's problem.

The cases differ in the nature of the plans adopted. In Cases 2 and 4, the ultimate plan was also a well-understood and frequently practiced pathway, although in Case 2 the new pathway contradicted the original. In case 3 delaying tactics and novel procedures were used while the group struggled to reach agreement; and in Case 1, the new plan was a novel application of a procedure typically used in other clinical situations, but neither used nor permitted in the current one. All 4 cases involved situations of significant risk, where unsuccessful outcomes would have led the decision-making to be strongly criticized, regardless of which plan was followed. In all cases, the change in plans was initiated by a single clinician but was only rapidly assented to by others involved in cases 1 and 4. Finally, all 4 cases, the change in plan did not originate from conscious, explicit reasoning about alternatives, but "came into mind" suddenly, apparently unbidden, but with a convincing sense that the change was in fact the correct course of action.

LIMITATIONS

The cases discussed here unfortunately do not cover the entirety of sudden changes in plan, as all turned out well. This undoubtedly represents a selection bias on at least the part of our informants, who could choose which cases they were willing to relate. We hope to elicit contrasting cases where sudden changes in plan resulted in going off course.

CONCLUSION

Naturalistic decision making research has highlighted how often experts' initial (or first few decision options) turn out to be satisficing if not optimal. These cases extend that thinking to situations in which the initial course has already been set, but dramatic changes are made without correspondingly dramatic changes in positive or negative cues. In these cases, the new option sprang to a practitioner's mind unbidden, and immediately "seemed right." Except for Case 2, they were not preceded by a general sense that something was not right; instead they thought they were on the right path until the intrusion of another, better and immediately convincing alternative persuaded them they were not.

ACKNOWLEDGMENTS

We thank all the clinicians who provided cases, details and insights to us.

REFERENCES

Brehmer, B. (1987). Development of mental models for decision in technological systems. In J. Rasmussen, K. Duncan & J. Leplat (Eds.), *New Technology and Human Error* (pp. 111 - 120). Chichester, UK: John Wiley & Sons.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. Cambridge, MA: The MIT Press.

Fundamental on Situational Surprise: A Case Study with Implications for Resilience

Robert L Wears[1,2] and L Kendall Webb[3]
[1,3]University of Florida, 655 W 8th Street, Jacksonville, Florida 32209, USA
wears@ufl.edu; kendall.webb@jax.ufl.edu
[2]Imperial College London, Praed Street, London W2 1NY, UK
r.wears@imperial.ac.uk

Abstract. Fundamental surprise is a challenge for resilience, since by definition it cannot be anticipated, and monitoring is limited by the lack of knowledge about what to target. It does, however, present opportunities for both responding and for learning. We describe an incident in which we use the co-occurrence of situational and fundamental surprise to reveal patterns about how adaptive capacity was used to meet challenges, and what barriers to learning were present. We note that temporal and cross-level factors played important roles in affecting the balance between situational and fundamental learning. Because the situational story of component failure developed first, it was difficult for the fundamental story of unknown, hidden hazards to supplant it. In addition, the situational story was easily grasped by all members of the organization, but the implications of the fundamental story were difficult for non-technical members, including senior leadership, to grasp. The responses at both the situational and fundamental level contain information about both specific vulnerabilities and general adaptive capacities in the organization.

1 INTRODUCTION

"Things that never happened before happen all the time." (Sagan, 1993)

Analyses of critical incidents often distinguish between situational and fundamental surprise (Lanir, 1986; David D. Woods, Dekker, Cook, Johannesen, & Sarter, 2010). Events characteristic of situational surprise might be temporally unexpected, but their occurrence and evolution are generally explicable and more importantly, compatible with the ideas generally held by actors in the system about how things work or fail, and the hazards that they face. Fundamental surprise, on the other hand, is astonishing, often inexplicable, and forces the abandonment of the broadly held notions of both how things work, and the nature of hazards that are confronted.

If we think of a system's resilience as its intrinsic ability "to adjust its functioning prior to, during, or following changes or disturbances, so that it can sustain required

operations under both expected and unexpected conditions" (Hollnagel, 2011), then it is clear that surprise creates unexpected demands that call for a resilient response.

Lanir (Lanir, 1986) has identified 4 characteristics that distinguish situational from fundamental surprise. Fundamental surprise refutes basic beliefs about 'how things work', while situational surprise is compatible with previous beliefs. Second, in fundamental surprise one cannot define in advance the issues for which one must be alert. Third, situational and fundamental surprise differ in the impact of information about the future. Situational surprise can be averted by such foresight, while advance information in fundamental surprise actually causes the surprise. And finally, learning from situational surprise seems easy, but learning from fundamental surprise is difficult.

Resilience is also characterized by 4 cardinal activities: monitoring, anticipating, responding, and learning. While effective management of situational surprise would typically involve all 4 of these activities, fundamental surprise clearly is a profound challenge for resilience, because one cannot monitor or anticipate items or events that are inconceivable before the fact. This leaves only responding and learning as the immediately available resilience activities in fundamental surprise, and explains in part why fundamental surprise is such a challenge to organisational performance. However, fundamental surprise does afford opportunities for deep learning, in particular the development of 'requisite imagination', an ability to picture the sorts of unexampled events that might befall (Adamski & Westrum, 2003)

We present a case study of the catastrophic failure of an information technology (IT) system in a healthcare delivery organisation, and of the organisation's response to it from the point of view of resilience. The failure itself involved a combination of both situational and fundamental surprise. As might be expected, the immediate response involved both adaptations of exploitation (*ie*, consuming buffers and margin for manoeuvre to maintain essential operations) and adaptations of exploration (*ie*, novel and radical reorganisations of the way work gets done) (March, 1991). Because fundamental surprise makes the disconnect between self-perception and reality undeniable, it affords the opportunity for a thorough-going reconstruction of views and assumptions about how things work. However, in this case the conflation of fundamental and situational surprise led to a classic *fundamental surprise error* – a re-interpretation of the problem in local and technical terms.

2 THE CASE

In this section we describe the events and the adaptations to the interpretations made of them, based on notes, formal reviews, and interviews during and after the incident.

2.1 Events

Shortly before midnight on a Monday evening, a large urban academic medical centre suffered a major information technology (IT) system crash which disabled virtually all IT functionality for the entire campus and its regional outpatient clinics (Wears, 2010). The outage persisted for 67 hours, and forced the cancelation of all elective procedures on Wednesday and Thursday, and diversion of ambulance traffic to other hospitals. (52 major procedures and numerous minor procedures; at least 70 incoming ambulance cases were diverted to other hospitals). There were 4 to 6 hour

delays in both ordering and obtaining laboratory and radiology studies, which severely impacted clinical work. The total direct cost (not including lost revenue from cancelled cases or diverted patients) was estimated at close to $4 million. As far as is known, no patients were injured and no previously stored data were lost.

The triggering event was a hardware failure in a network component. This interacted with modules not known to be present but left behind from an incompletely aborted (and ironically named) "high availability computing" project some years previous; this interaction prevented the system from restarting once the network component was replaced. The restart failure could not be corrected initially because of a second, independent hardware failure in an exception processer. Once this was identified and replaced, the system still could not be restarted because unbeknownst to the IT staff, the permissions controlling the start-up files and scripts had been changed during the same project, so that no one in IT was able to correct them and thus re-start the system.

2.2 Adaptations

After a brief initial delay, the hospital was able to quickly reorganize in multiple ways to keep essential services operating for the duration. Adaptations included exploitation of existing resources or buffers; and exploration of novel, untried ways of working (March, 1991). These adaptations correspond roughly to the first- and second-order resilient *responses* described by a well-known materials science analogue (David D Woods & Wreathall, 2008).

Adaptations of exploitation included deferring elective procedures and speeding discharges of appropriately improving inpatients. The former was limited in scope because the extent of the problem was not realized until Tuesday's elective cases were well underway. The latter was stymied by the slow delivery of laboratory and imaging results; physicians were reluctant to discharge patients when results were still pending. This, of course, is one of the classic patterns of failure – falling behind the tempo of operations (David D Woods & Branlat, 2011).

Several adaptations of exploration were invoked. An incident command team was formed. Because the geographic area experiences frequent hurricanes, the incident command system was well-rehearsed and familiar, so it was adapted to manage a different type of threat.

A similar novel use of available techniques evolved dynamically to compensate for the loss of medical record numbers (MRNs) to track patients, orders, and results while the system was down. The emergency department had been planning to implement a 'quick registration' method, where only basic patient information is obtained initially to permit earlier orders and treatment, and the registration process is completed at a later time. The IT failure prevented complete registration but was thought to have left the capability for quick registration. This method was very close to implementation, so it was pressed into service. However, its application in this setting uncovered a problem, in that different organisational units used the same variable to represent different information; this resulted in several patients in getting "lost" in the system. This failure led to an alternative, the use of the mass casualty incident (MCI) system.

In many MCIs, the numbers of arriving patients would too rapidly exceed the ability to record their basic information and assign them identifying MRNs, so the organization maintained a separate system with reserved MCI-MRNs and pre-printed arm-

bands. Although this system was envisioned for use in high demand situations, it was formally designed to accommodate any mismatch between demand and available resources. In this case, demand was normal to low, but resources were much lower, so the MCI system was used to identify and track patients and marry them to formal MRNs after the incident had been resolved.

The most novel adaptation of exploration included rescheduling financial staff (who now had nothing to do, since no bills could be produced or charges recorded) as runners to move orders, materials, and results around the organization that had previously been transmitted electronically.

2.3 Interpretations

The case was viewed in multiple ways within the organisation, depending on the orientation to situational or fundamental surprise. It should be emphasized that there is not a 'correct' interpretation here – both views have both validity and utility, and both must be understood and held simultaneously for a full understanding of the case and its implications for organisational resilience.

Situational Surprise. Because the triggering event was a hardware failure, and because the organisation had experienced a similar incident leading to total IT failure 13 years previously (Wears, Cook, & Perry, 2006), the failure was initially interpreted as a situational surprise. It evinced no fundamental misperception of the world; it was not 'the tip of the iceberg' but rather a hazard about whose possibility there had always been some awareness.

However, we should not downplay the importance of the organisation's situational response, which was in many ways remarkably good. The organisation detected the fault and *responded* relatively quickly and effectively; the unfolding understanding of the situation and effectiveness of the response was *monitored*, and the organisation reconfigured to meet the threat. This reconfiguration involved a mixed control architecture where a central, incident command group set overall goals and made global level decisions (*eg*, cancelling elective procedures, reassigning financial staff) and managed communications among the various subunits of the organisation, while allowing functional units (*eg*, the emergency department, operating room, intensive care units, pharmacy, radiology, laboratory, and nursing) to employ a mixture of pre-planned and spontaneously developed adaptations to maintain performance.

There was a specific attempt to capture situational *learning* from the incident. Each major unit conducted its own after action review to identify performance issues; the incident command group then assembled those and conducted a final, overall review to consolidate the lessons learned. This review obtained broad participation; it resulted in 104 unique items that, while locally oriented and technically specific, form the nidus of organisational memory and could inform the approach to similar future events, which are broadly *anticipated* in their consequences (*ie*, another widespread IT failure at some point seems assured) if not in their causes.

One remarkable aspect of the response was the general absence of finger-pointing or accusatory behaviours, witch-hunts or sacrificial firings. An essay on how complex systems fail (Cook, 2010) had been circulated among the senior leaders and the incident command group during the outage, with substantial agreement on how well it described the incident, its origins, and consequences.

Fundamental Surprise. However, as a fuller understanding of the incident developed, situational gave way to fundamental surprise. The discovery of the permissions problem refuted taken-for-granted beliefs – that IT services understood and could maintain their own systems; and in particular, that restrictions to privileged ("root") access could not be compromised except by sabotage. It raised the question of what other, previously unknown threats, installed by a parade of vendors and consultants over the years, lay lurking just beneath the surface waiting to be triggered into behaviours both unexpected and unexplainable.

Lanir notes that "when fundamental surprises emerge through situational ones, the relation between the two is similar to that between peeled plaster and the exposed cracks in the wall. The plaster that fell enables us to see the cracks, although it does not explain their creation" (Lanir, 1986). The IT unit recognized this clearly, and were astonished by the "hidden time bomb" whose presence was only fortuitously revealed by the line card failure. This triggered a deeper review of known previous changes, a new commitment to not permitting unmonitored and undocumented changes by vendors or other 3rd parties, and more stringent requirements for "as installed" documentation (including personal identification of involved parties). It led to a general awareness among IT leaders that their knowledge of their own system was incomplete and that they should therefore act in the "continuing expectation of future surprise" (Rochlin, 1999). This fundamental learning, however, did not spread throughout the organisation, but remained mostly encapsulated in IT.

3 DISCUSSION

Critical incidents are ambiguous: is stopping short of complete breakdown a story of success, or a harbinger of future failure (David D Woods & Cook, 2006)? Incidents embody a dialectic between resilient adaptation and brittle breakdown. In this case we see successful, resilient adaptation, but the real lesson is not in the success but rather in how adaptive capacity was used, and how it can be fostered and maintained. We also see limited fundamental learning, but the real lesson is not the failure of more broadly based learning but rather understanding what made that learning difficult.

3.1 Fundamental surprise as a challenge to resilience

Fundamental surprise represents a major challenge to organisational resilience. Since by definition, fundamental surprise events are inconceivable before the fact, they cannot be anticipated; since it is unknown whence the come, there can be little guidance on what, exactly, to monitor to facilitate their detection.

3.2 Factors limiting fundamental learning

There is a strong tendency to re-interpret fundamental surprise in situational terms (Lanir, 1986). Several factors combined to limit fundamental learning in this case.

Situational Surprise. The co-occurrence of a situational surprise (failure secondary to component failure) made it easy to redefine the issues in terms of local technical problems (*eg*, the lack of available spares). The easy availability of hardware failure as an explanation for the outage limited deeper analysis and understanding. In addition, the relative success of the adaptations to the failure paradoxically made deeper understanding seem less important.

Temporal Factors. The full understanding of the incident did not develop until roughly 36 hours into the outage, so the initial characterisation of the problem as a hardware issue proved hard to dispel. In addition, the 24 x 7 x 365 nature of healthcare operations required urgent responses to prevent immediate harm to patients. This narrowed the focus of attention to actions that could be taken immediately to manage the disturbance, and moved deeper understanding to a lower priority.

Cross-level Interactions. Different understandings were held at different levels of the organisation. The technical problem – unauthorized, unrecognized access to critical files – was harder for non-technical leadership to understand, particularly compared to the easily grasped story of component failure. Although one might suspect that the full story might have been embarrassing thus obscured or suppressed, this was not the case; the IT leadership was remarkably forthcoming in laying out the full explanation of what was known, as it became known.

In addition, one might question whether it was even pertinent for the clinical arm of the organisation to undergo fundamental learning. Clinical operational units need to be prepared for the consequences of IT failures, but have little role in anticipating or preventing them.

Healthcare-specific Factors. IT in healthcare has several unique characteristics that contributed to both the incident and to the difficulty of fundamental learning. In contrast to other hazardous activities, IT in health is subject to no safety oversight whatsoever. The principles of safety-critical computing are virtually unmentioned in a large medical informatics literature (Wears & Leveson, 2008). Thus there is no locus in the organisation responsible for the safety of IT, and no individual or group who might be responsible for deeper learning from the incident.

In addition, IT in healthcare is relatively new compared to other industries. The systems in use today are fundamentally "accidental systems", built for one purpose (billing), and grown by accretion to support other functions for which they were never properly designed. This has led to "criticality creep", in which functions originally thought to be optional gradually come to be used in mission-critical contexts, in which properties that were benign in their original setting now become hazardous (Jackson, Thomas, & Millett, 2007).

Diverting Factors. Finally, an external factor diverted at least senior leadership's attention from a deeper exploration of the vulnerabilities whose presence this incident suggested. Nine months prior to this incident, the larger hospital system of which this organisation is a part made a commitment to install a monolithic, electronic medical records, order entry and results reporting system, provided by a different vendor across the entire system. Although full implementation was planned over a 5 year span, major components of the new system were scheduled to go live 9 months after the incident. This project gave the (misleading) appearance of a clean replacement of the previous system, a *deus ex machina,* and thus limited the felt need to understand the vagaries of the existing system more deeply, in addition to consuming a great deal of discretionary energy and resources.

CONCLUSION

Fundamental surprise is a challenge for organisational resilience because anticipation is not a factor and monitoring is limited, typically, to evaluating the quality of response. Fundamental surprise also affords great opportunities for deep and fun-

damental learning, but it is difficult to effectively engage organisations fully in the learning process. In this case, the combination of situation and fundamental surprise blurred the distinction between them; situational adaptation and learning were remarkable, but the ease of re-interpreting fundamental as situational surprise meant fundamental learning was encapsulated, limited to only parts of the organisation.

REFERENCES

Adamski, A. J., & Westrum, R. (2003). Requisite imagination: the fine art of anticipating what might go wrong. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (Vol. 193 - 220). Mahwah, NJ: Lawrence Erlbaum Associates.

Cook, R. I. (2010, 2010). How Complex Systems Fail Retrieved 19 September 2010, from http://www.ctlab.org/documents/Ch%2007.pdf

Hollnagel, E. (2011). Prologue: the scope of resilience engineering. In E. Hollnagel, J. Pariès, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice: A Guidebook* (pp. xxix - xxxiv). Farnham, UK: Ashgate.

Jackson, D., Thomas, M., & Millett, L. I. (Eds.). (2007). *Software for Dependable Systems: Sufficient Evidence?* Washington, DC: National Academy Press.

Lanir, Z. (1986). Fundamental Surprises Retrieved from http://csel.eng.ohio-state.edu/courses/ise817/papers/Fundamental_Surprise1_final_copy.pdf

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71 - 87.

Rochlin, G. I. (1999). Safe operation as a social construct. *Ergonomics, 42*(11), 1549 - 1560.

Sagan, S. D. (1993). *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton, NJ: Princeton University Press.

Wears, R. L. (2010). Health information technology risks. *The Risks Digest, 26*(25). Retrieved from http://catless.ncl.ac.uk/Risks/26.25.html#subj1

Wears, R. L., Cook, R. I., & Perry, S. J. (2006). Automation, interaction, complexity, and failure: a case study. *Reliability Engineering and System Safety, 91*(12), 1494 - 1501. doi: 10.1016/j.ress.2006.01.009

Wears, R. L., & Leveson, N. G. (2008). "Safeware": safety-critical computing and healthcare information technology. In H. K, J. B. Battles, M. A. Keyes & M. L. Grady (Eds.), *Advances in Patient Safety: New Directions and Alternative Approaches.* (AHRQ Publication No. 08-0034-4 ed., Vol. Vol 4. Technology and Medication Safety, pp. 1 - 10). Rockville, MD: Agency for Healthcare Research and Quality.

Woods, D. D., & Branlat, M. (2011). Basic patterns in how adaptive systems fail. In E. Hollnagel, J. Paries, D. D. Woods & J. Wreathall (Eds.), *Resilience Engineering in Practice* (pp. 127 - 144). Farnham, UK: Ashgate.

Woods, D. D., & Cook, R. I. (2006). Incidents -- markers of resilience or brittleness? In E. Hollnagel, D. D. Woods & N. Levenson (Eds.), *Resilience Engineering* (pp. 70 - 76). Aldershot, UK: Ashgate.

Woods, D. D., Dekker, S., Cook, R., Johannesen, L., & Sarter, N. (2010). *Behind Human Error* (2nd ed.). Farnham, UK: Ashgate.

Woods, D. D., & Wreathall, J. (2008). Stress-Strain Plots as a Basis for Assessing System Resilience. In E. Hollnagel, C. P. Nemeth & S. W. A. Dekker (Eds.), *Resilience Engineering: Remaining Sensitive to the Possibility of Failure* (pp. 143 - 158). Aldershot, UK: Ashgate.

Emergency Department Crowding:

Vicious Cycles in the ED

June 2011

J. Bradley Morrison, PhD
MIT Sloan School of Management
MIT Engineering Systems Division
Brandeis International Business School
morrison@mit.edu
(781) 736-2246

Robert L. Wears, MD, MS
University of Florida
Imperial College London
École des Mines de Paris

We are grateful for support provided by NIH Grant 1 R21 HL098875, the National Heart Lung & Blood Institute, and the Office of Behavioral and Social Science Research

Emergency Department Crowding: Vicious Cycles in the ED

Morrison and Wears

Abstract

Over the past several decades, demands on the United States emergency and trauma care system have grown dramatically, but the capacity of the system has not kept pace. The result is a widespread phenomenon of crowded emergency rooms, especially in urban hospitals, which has become a major barrier to receiving timely care and has been implicated in adverse medical outcomes. This paper develops a stylized system dynamics model to examine the dynamics of patient flow in emergency departments. Simulation results show that increased ED resilience can come from relaxing bed constraints or from more human capability to cope with increasing workloads. The vulnerability of this system is rooted in the critical interaction between physical constraints imposed by the environment and the human capability of the staff to work at high performance levels under conditions of worsening workload pressure.

Emergency Department Crowding: Vicious Cycles in the ED

Introduction

ED / hospital crowding is an international problem, affecting hospitals throughout the English-speaking world. The problem first became apparent in US EDs in the 1980s, and was thought to be of crisis proportions by the end of that decade. The American College of Emergency Physicians issued a position statement (American College of Emergency Physicians 1990) and several policy recommendations (American College of Emergency Physicians 1990) at was then called "emergency department overcrowding" in 1990, but the problem only continued to grow (Derlet and Richards 2000; Goldberg 2000; Kellermann 2000; Zwemer 2000). Eleven years later, in 2001, the Society for Academic Emergency Medicine (SAEM) made crowding the theme of its yearly Consensus Conference; entitled *The Unraveling Safety Net*, the Conference resulted in the dedication of an entire issue of the Society's journal, *Academic Emergency Medicine*, to a group of papers on the crowding problem (Adams and Biros 2001; Baer, Pasternack et al. 2001; Derlet, Richards et al. 2001; Gordon, Billings et al. 2001; Kelen, Scheulen et al. 2001; Reeder and Garrison 2001; Schneider, Zwemer et al. 2001; Schull, Szalai et al. 2001). Despite this attention, crowding has only gotten worse in the ensuing years (US General Accounting Office 2003; Kellermann 2006), culminating in a 2006 Institute of Medicine report that warned that the system was on the verge of total breakdown (Institute of Medicine 2006); despite this attention, and a plethora of interventions aimed at mitigating it, crowding seems to have been monotonically increasing over the past 25 years or so.

There have been multiple attempts to develop a workable definition of crowding (Hwang and Concato 2004). A recent systematic review of the crowding literature (Hoot and Aronsky 2008) concluded that the American College of Emergency Physician's consensus definition seemed to encompass most of the important and relevant aspects of the problem: "Crowding occurs when the identified need for emergency services exceeds available resources for patient care in the

emergency department, hospital, or both." (American College of Emergency Physicians 2006) This definition highlights crowding as an imbalance between supply and demand, and, as modified by Pines to include an impact on the quality of care (Pines 2007), has been widely accepted among researchers. Asplin *et al* (Asplin, Magid et al. 2003) advanced the understanding of the crowding problem by developing a conceptual model that provided a practical and now widely accepted framework for research, policy and management addressing crowding. The model (see Figure 1) partitions the problem space into 3 interacting components: input, throughput, and output, and has become generally accepted in healthcare in discussions of the crowding issue. Input factors reflect the sources and aspects of patient inflow; throughput factors reflect bottlenecks and delays within the ED; and output factors reflect bottlenecks in other parts of the healthcare system that might affect the ED.

Crowding has multiple, complex, interacting causes, and many 'obvious' causes have been discredited (Derlet and Richards 2000). Roughly 1/3 of the papers Hoot and Aronsky (Hoot and Aronsky 2008) included in their systematic review concerned research into the causes of crowding. These works tend to naturally fall into two separate areas, one concerned with general, long term trends and conditions, and the other with more specific, often local, triggering factors.

The long term trends are summarized by growing demand and falling supply. From 1995 to 2005, annual ED visits increased by 20% (from 96 to 115 million) and per capita ED visits by 7% (from 37 to 40 visits per 100) (Nawar, Niska et al. 2007). During the same period, the number of EDs decreased by 381, the number hospitals decreased by 535, and the number of hospital beds by 134,000 (Nawar, Niska et al. 2007; Health Forum 2008). In this view, crowding (and its consequences) is the inexorable result of long-term secular trends.

While not denying the influence of these general causal factors, work on specific factors has addressed issues such as ED use for non-urgent problems, by the uninsured, or by frequent users; and issues related to internal ED operating efficiency.

Work on crowding was initially held back by a number of assumptions, or "folk models" about its causes that ultimately proved to be false, or at least misleading (Newton, Keirns et al. 2008). For example, it has been widely thought that ED crowding is due to increased numbers of patients with relatively trivial, non-emergent problems, to increasing numbers of uninsured patients, or to "frequent flyers" – repeat visits by a small number of patients (Washington, Stevens et al. 2002). None of these hypotheses have been substantiated, and there is countervailing evidence for each (Sprivulis, Grainger et al. 2005). For example, Schull *et al* (Schull, Kiss et al. 2007) studied 110 EDs and 4.1 million patient visits in Ontario, and found that low-complexity patients contributed only trivially to length of stay and physician treatment times (32 and 13 seconds per patient, respectively). The same group also showed that ambulance diversion was not associated with either low complexity patients or with throughput factors, but was associated with output factors (Schull, Lazier et al. 2003). The results were similar across moderate and high volume EDs, and were robust to variations in the definition of low complexity. These results suggest that attempts to divert low-complexity patients to alternative sources of care are unlikely to substantially improve ED flow or to alleviate ED crowding. While this study does not dismiss the concern about nonurgent ED use as a policy issue – patients should not be forced into using the ED because they have no alternative –it does show that diverting low urgency patients away from the ED will not have a significant impact on crowding.

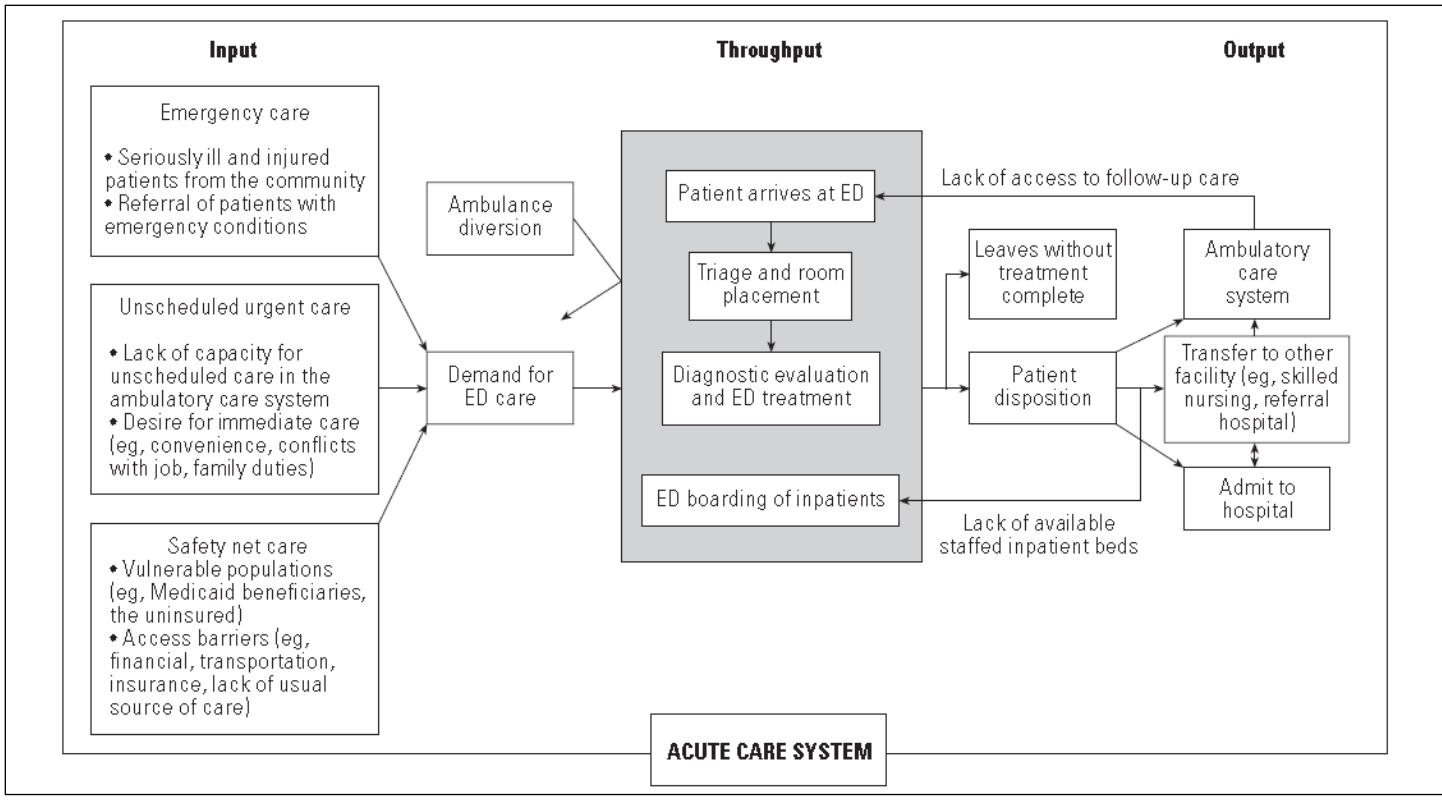


Figure 1. The input-throughput-output conceptual model of crowding (Asplin, Magid et al. 2003).

The problem was first framed as "ED crowding", and initial work considered the ED in isolation – input and output factors were considered uncontrollable or at least outside the scope of ED managers who were dealing with the problem; in addition, 20 years ago, many ED inefficiencies did exist. However, as these inefficiencies were gradually wrung out of ED systems of care, the potential for alleviating crowding by addressing throughput issues has diminshed. The weight of recent research has led to the conclusion that "… ED crowding is a local manifestation of a systemic disease" (Hoot and Aronsky 2008), and that effective solutions will have to set a scope that includes both input and output factors (Litvak, Long et al. 2001; Forster, Stiell et al. 2003; Richardson 2003). For example, systematic hospital restructuring has been shown to lead to subsequent crowding (Schull, Szalai et al. 2001). In another study, Rathlev *et al* (Rathlev, Chessare et al. 2007) retrospectively analyzed 93,000 visits at a single academic ED to describe the association of various input, throughput, and output factors on ED

length of stay. The only factors that were associated with increased length of stay were output factors: hospital occupancy, number of ED admissions to the hospital, and number of elective surgical admissions. The organizations that have had the greatest success in managing crowding have been those that recognized the hospital-wide nature of the patient flow problem and designed initiatives to address ED output at the organizational level (Cardin, Afilalo et al. 2003; Asplin and Magid 2007).

ED / hospital crowding leads to poorer outcomes in a variety of important conditions and patient groups, in brief, it hurts patients and degrades the quality of care (Bagust, Place et al. 1999; Richardson 2006; Sprivulis, Da Silva et al. 2006; Weissman, Rothschild et al. 2007). Crowding has been associated with delays in treatment (JCAHO 2002), increases in inpatient length of stay (Richardson 2002), particularly in the elderly (Liew and Kennedy 2003) and with increased mortality in hospitalized patients (Richardson 2006; Sprivulis, Da Silva et al. 2006) . One of the earliest symptoms of crowding was the problem of ambulance diversion (Goldberg 2000; Eckstein, Isaacs et al. 2005; Burt and McCaig 2006; Sprivulis, Da Silva et al. 2006). Crowding has been associated with lower quality care for chest pain patients (Diercks, Roe et al. 2007), and delays in ED care (Schull, Morrison et al. 2003) and in delivery of definitive care such as fibrinolysis or catheterization in acute myocardial infarction (Schull, Vermeulen et al. 2004), and in worsened cardiac outcomes (Pines and Hollander 2007). It is associated with delays in antibiotic administration in serious infections (Fee, Weber et al. 2007; Gray and Baraff 2007; Pines, Localio et al. 2007) and deficient pain management (Hwang, Richardson et al. 2006) in the ED. In hospital care, crowding is associated with increases in adverse events (Cameron 2006), and in premature discharges from inpatient care (Baer, Pasternack et al. 2001; Jack, Chetty et al. 2009). Virtually every group of patients have been affected, but vulnerable populations, such as children (Committee on Pediatric Emergency Medicine 2004; Lorch, Millman et al. 2008) or the elderly are particularly susceptible (Hwang, Richardson et al. 2006).

ED – hospital crowding has shown "policy resistance" and has resisted efforts to alleviate or mitigate it. One of the striking observations about the ED-hospital crowding problem is its

persistence despite general agreement that it hurts both patients and health care organizations (Bagust, Place et al. 1999; Bayley, Schwartz et al. 2005; Falvo, Grove et al. 2007; Falvo, Grove et al. 2007). Multiple authors have raised the question of why it persists and in fact has worsened, in the face of multi-faceted attempts to control it (Kellermann 2000; Agrawal 2007; Kelen and Scheulen 2007; Moskop, Sklar et al. 2008; Moskop, Sklar et al. 2008; Viccellio 2008). This seems to be a classic case of "policy resistance", arising, as Sterman (Sterman 2000) has suggested, from an incomplete understanding of the problem; essentially, researchers have been "looking in the wrong place" for insights into the crowding problem (Lane, Monefeldt et al. 2000).

Crowding exhibits many of the characteristics that are best addressed in a system dynamics approach. It shows non-linear dynamics analogous to phase shifts in physics (Hollnagel and Sundström 2006; Wears and Perry 2006; Woods, Wreathall et al. 2006), punctuated equilibria in biology (Gould 1989), or domain shifts in ecology (Holling 1973; Lesne 2008). Hwang and Lichtenthal's characterization of slowly developing organizational crises seems apt here (Hwang and Lichtenthal 2000). In this paradigm, a slow change in a critical variable, which may be well known and easily identified, leads to a relatively sudden and discontinuous change in the behavior of the system when a threshold value is crossed; this is often accompanied by hysteresis – although a small increment in the critical variable may have led to a large change in the system, a subsequent small decrement will not restore the system to its previous state (Anderies, Walker et al. 2006; Walker and Salt 2006).

In addition, crowding shows delayed feedback loops (Hollander and Pines 2007) and complex interactivity. "Access block" – the inability to move admitted patients out of the ED because no inpatient beds are available – is associated with increased length of stay in hospitalized patients, which of course makes crowding and access block worse (Richardson 2002; Forster, Stiell et al. 2003; Liew and Kennedy 2003). Attempts to alleviate crowding often place pressure on physicians to discharge patients from the hospital sooner, but premature discharges lead to an increase in return visits to the ED by patients who are more complex, tend to stay longer, and are more often re-admitted (Baer, Pasternack et al. 2001; Jack, Chetty et al. 2009).

Many of the proposed interventions for crowding offer temporary respite but are either unsustainable or in the long run counterproductive. Where inpatient capacity is truly inadequate, increasing the supply of inpatient beds is of course indicated, but as a general solution is clearly unsustainable. Improving ED throughput by increasing departmental efficiency has been a central focus of effort, but recent studies of crowding have shown that both input and throughput factors are not associated with crowding, whereas output factors were (Rathlev, Chessare et al. 2007). Essentially, it seems that throughput factors have been optimized already, because the ED managers have been closest to the problem for many years and these factors are within their span of control; thus there is little further to be gained by incremental increases in ED efficiency (Karpiel 2004; King, Shaw et al. 2004; Patel, Derlet et al. 2006; Shah, Fairbanks et al. 2006; Worster, Fernandes et al. 2006). Other popular solutions, such as moving "boarded" patients from ED hallways to hallways on inpatient wards (Viccellio 2001), simply shift the location of the problem without addressing it in a fundamental way. Similarly, ambulance diversion has been shown to shift crowding from one hospital to another, and sometime to trigger a series of 'tit-for-tat' diversions that simply further increase congestion in the system (Asamoah, Weiss et al. 2008).

A final, minimal approach to the problem has been to manage it by fiat. The Joint Commission has declared ED – hospital crowding unacceptable, and that organizational leadership should "… develop and implement plans to identify and mitigate … overcrowding" (Joint Commission on Accreditation of Healthcare Organizations 2003) without notable effect. In the UK, crowding became a *cause célèbre* and led to a "4 hour mandate" – an NHS regulation that patients in the ED must be either admitted, transferred or discharged within 4 hours of the time they first signed in to the department (Department of Health 2000), enforced by financial sanctions on the organization for breaches. An analysis of the effect of this mandate shows a shifting of the problem – a sharp peak in hospital admissions and ED discharges just at 4 hours (Locker and Mason 2005). One of the effects of the 4 hour mandate in UK hospitals has been that the majority of these "admissions" are to a unit which is another part of the ED in all but name,

satisfying the technical requirements of the rule but having less effect on the problem (Weber, Mason et al. 2011).

Because of its dynamic complexity, delayed feedback loops, and social-behavioral components, the problem seems ideally suited to a system dynamics approach (Homer and Hirsch 2006), but it has been infrequently used. A Pubmed search for the terms 'system dynamics' and 'emergency' in any text field yielded only 6 citations, but only 2 of these were directly relevant. (By comparison, a search for 'pancreatitis' yields almost 45,000 citations). One of these studies was narrowly focused on laboratory response time and its effect on ambulance diversion (a proxy for crowding) (Storrow, Zhou et al. 2008); it showed a strong association between laboratory turnaround time and several measures of ED efficiency. The other (Lattimer, Brailsford et al. 2004) examined ED use at a regional rather than an organizational level, and predicted that ED volumes would increase, leading to increases in hospital occupancy and eventually "bottlenecks" – *ie*, crowding – in the region. One additional paper not listed in Pubmed focused primarily on the tradeoff between beds for emergency admissions and those for elective surgery admissions, but not on the origins and persistence of crowding itself (Lane, Monefeldt et al. 2000).

Several other approaches have been explored, including discrete event simulation (Bagust, Place et al. 1999; Hoot, LeBlanc et al. 2008), queuing theory (Litvak, Long et al. 2001; Litvak, Buerhaus et al. 2005), and other engineering methods (Levin, Han et al. 2007; Levin, Dittus et al. 2008). While these approaches have provided useful insights, they have not addressed the central issue of whether the structure of the system itself produces the phenomenon of crowding.

Therefore, the broad, overall objective of this paper is to use system dynamics modeling (Sterman 2000) to study the problem of emergency department (ED) and hospital crowding in order to inform departmental, organizational, regional, and societal policies and interventions

aimed at alleviating it. For example, a system dynamics understanding of crowding would be useful in the following ways:

- Developing early warning capabilities of a potential overcrowding crisis
- Identifying leverage points for managing dynamic and unexpected changes in patient demand or organizational capacity to respond
- Identifying potentially dysfunctional interventions to be avoided, *ie*, that might provide short term relief but ultimately make the overall problem worse.

The model development and analysis that follow are motivated by ethnographic observation of the day-to-day operating practices in the emergency department, including a level 1 trauma center, of a large, inner-city teaching hospital and by one author's first-hand experience as an emergency physician. The paper draws on data sources (not presented here) comprising observations, interviews, archival data, and the literatures in medicine, health care, the management sciences and organizational theory to inform the development of a system dynamics model and analysis that explores the phenomenon of emergency room crowding, with a particular focus on how the people and systems on the front lines adapt and adjust to cope with the challenges of excess demand.

Model Development

The input-throughput-output framework shown in Figure 1 is the starting point for our model development (Asplin, Magid et al. 2003). We begin by carefully distinguishing the stocks and the flows. Stocks are accumulations, such as the accumulation of patients in the ED. Flows cause increases or decreases in stocks. The framework depicts three sources of inputs that generate demand for ED care, which is the inflow to the stock of patients in the ED. The figure also shows two paths by which patients exit the ED, which are outflows from the stock of patients in the ED. Thus, "patient disposition" and "leave without treatment complete" are two outflows from the stock. The outflow labeled patient disposition comprises three possibilities - admit, transfer, or discharge to the ambulatory

care system. Finally, the figure also shows that patients returning from the ambulatory care system constitute another inflow to the stock of patients in the ED.

Figure 2 uses the traditional icons for system dynamics models to depict the stock and flow structure of this system. Stocks are represented by rectangles. Flows are represented by the pipe and valve icons. Each stock and flow is labeled with a variable name.

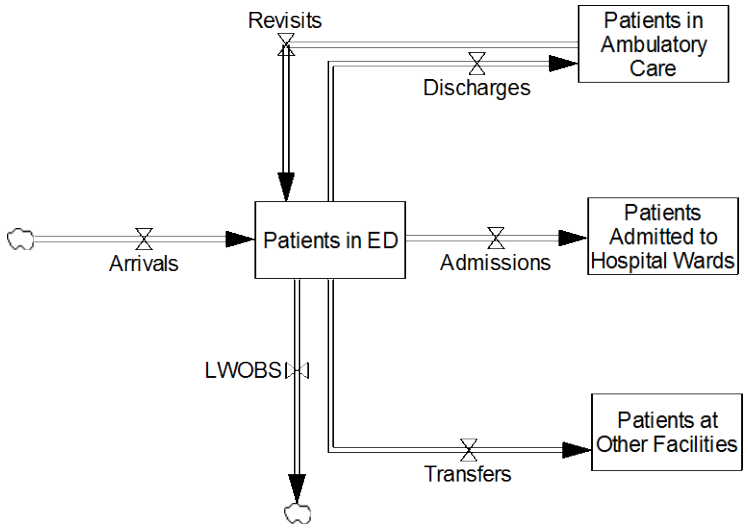


Figure 2. The stock and flow structure of the input-throughput-output framework. Stocks are depicted by rectangles. Flows are depicted by the pipes and valves. Clouds represents sources and sinks that are considered outside the model boundary.

The aim of the remainder of this paper is to develop and analyze a conceptual model of patient flows that allows us to examine some, but perhaps not all, meaningful aspects of the dynamics of ED crowding. The modeling process is iterative, and the choice of what to include in a model is based on the purpose of the model (Randers 1980; Homer 1996) . Our purpose here is to begin to understand how patient management practices in the ED interact with elements of the broader health care system within which the ED functions, so we have chosen to include one aspect of patient management - decision making for patient disposition - and one aspect of the hospital system - admission to the wards.

We present the model here in stages, beginning with a model that focuses on the physical movement of patients, expanding on the structure shown in Figure 2. We turn our attention first to the admission process. When an ED physician (or physician team) decides that the proper disposition for a patient is to be admitted to the hospital wards, the decision triggers a complex process that usually leads to the physical transfer of the patient from the ED to the hospital ward. The ED issues a request for a consultation from a relevant specialist or general practitioner with admitting privileges. If the consulting physician concurs with the ED physician's recommendation to admit the patient, the consulting physician writes admitting orders, initiating a request for assigning a bed to this patient. Once the patient has a bed assigned, the transport personnel in the hospital may physically move the patient to the hospital ward. The structure shown in Figure 3 adds the stocks and flows describing these key steps. The large rectangle around the three stocks of patients Awaiting Consults, Awaiting Assigns, and Awaiting Transport signals that these patients are typically still physically located in the ED. (For the purpose of this early conceptual model of the dynamics of ED patient flow, we will ignore the outflows for LWOBS and Transfers shown in Figure 2.)

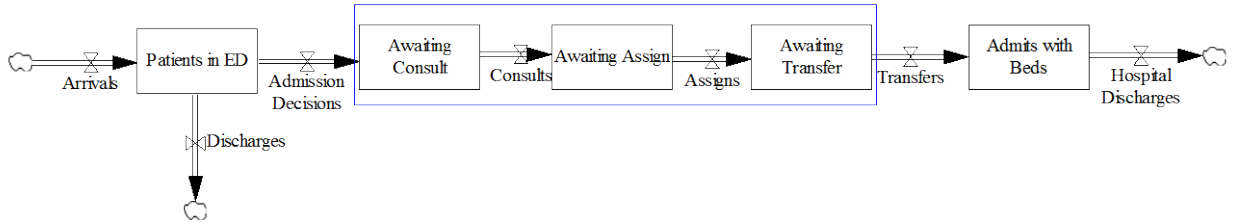


Figure 3. The stock and flow structure with detail on hospital admissions.

The rates of patient flows will depend on various factors, including factors based on waiting times and processing times, available resources, and other capacity constraints. The available time for consulting physician specialists is an example of a capacity constraint that can affect the rate of Consults. The time required for a consulting physician to become free and to travel to the ED to see a patient contributes to waiting time. The time for communicating with the ED physician and evaluating the patient constitute processing time. Similarly, there are various

activities and delays associated with Assigns and Transports. We model these flows of (Consults, Assigns, Transports) by assuming an average elapsed time that comprises the waiting and processing times and further assume that this average time is constant. We also explicitly model how constrained availability of beds when hospital occupancy is high affects the rate of flow of Assigns. Because there is a fixed number of beds in the hospital, when the hospital approaches full occupancy, it becomes increasingly difficult to assign a bed to a patient. The rate of inflow to the stock of Admits with Beds must slow down, and indeed if the hospital is completely full must equal zero. The model captures this critical feedback process explicitly, as shown in Figure 4. The rate of Assigns is the lesser of the Desired Rate of Assigns and the Feasible Rate of Assigns. The Desired Rate of Assigns is a constant fraction per unit time of the stock of patients Awaiting Assign, representing the demand for beds from patients ready to be assigned. The Feasible Rate of Assigns represents the supply of beds that can be assigned to these patients. Beds may be available because there are empty beds (i.e., occupancy is less than 100%) and because patients get discharged, freeing their beds for reassignment. Thus, the Feasible Assignment Rate is the sum of the rate of assigning previously empty beds such that occupancy increases and the rate at which beds become available from Hospital Discharges. In most real hospitals, patients from the ED are only one source of demand for hospital beds. Others include surgical admissions and medical admissions directly from other specialties. The model here does not include other sources of demand, the bed capacity to serve them, or the decision making processes for assigning beds to these competing sources of demand. Instead, we interpret the fixed quantity of beds in the model as representing the beds allocated to patients from the ED.

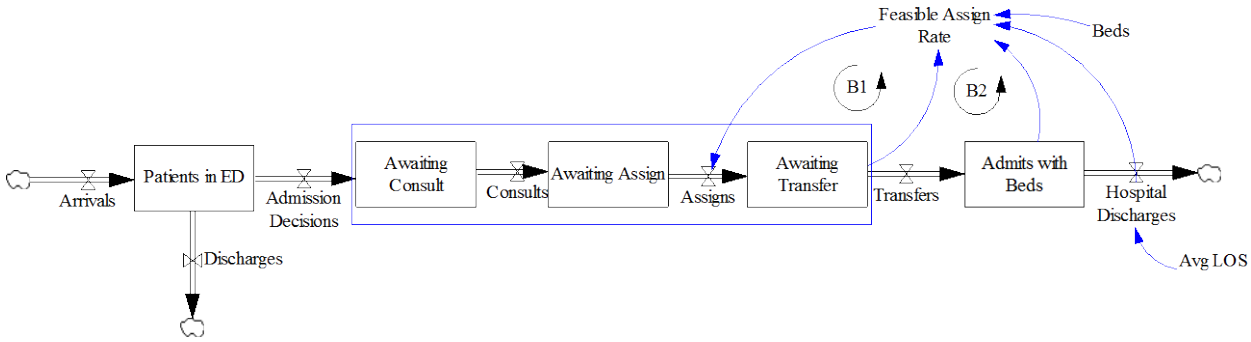


Figure 4. The feedback structure of the constraint imposed by hospital bed availability.

In Figure 4, the lines with arrows are causal links. A causal link from one variable to another variable (which can be a flow) means that a change in the first variable causes a change in the second variable. For example, an increase in the rate of Hospital Discharges causes an increase in the Feasible Assign Rate. Conversely, an increase in the number of Admits with Beds causes a decrease in the Feasible Assign Rate, because the number of empty beds is lower. Together with the stocks and flows, the causal links form feedback loops. For example, imagine the stock of Admits with Beds increases (due to an inflow of Transfers). The increase in Admits with Beds causes a decrease in the Feasible Assign Rate. As this rate falls low enough, it causes the rate of Assigns to decrease. As the inflow of Assigns drops below the outflow of Transfers, the stock of Awaiting Transfers decreases, which reduces the rate of Transfers, slowing or stopping the increase in stock of Patients with Beds. The feedback process works to offset, or balance, the original change (the increase in Patients with Beds), so we designate this a balancing loop. Two such loops are labeled in Figure 4 as B1 and B2. Balancing loops bring stability to systems, often by limiting growth or moving the system towards some implied target. In this case, the loops act as controls on the inflow of patients to the wards given the physical reality that a bed must be available in order to assign a bed.

To use this model to investigate the dynamics of patient flow, we specify equations for each variable shown in the diagram. Appendix 1 presents the full equation listing. The equations translate the causal logic shown in the diagram into algebraic representations. Parameter values are required for constants such as average time delays (e.g., Avg LOS) and number of Beds. For our conceptual analysis here, we use parameter values suggested by practicing emergency physicians. Arrivals to the ED tend to be lowest in the early morning hours, rise to a peak in the late afternoon (around 4:00 or 5:00 pm) and then taper off throughout the night. The simulations in this paper all begin with an arrival flow that mimics this diurnal cycle as shown in Figure 5 generated by an average arrival rate adjusted by a diurnal multiplier. Discharges from the hospital are also subject to some of the same diurnal factors, so we adjust the endogenously generated rate of discharges by the same diurnal multiplier. We set the initial

conditions for all stocks to the long-term steady state values for midnight (because time 0 is midnight of the first the day) so the model begins near a steady-state. Figure 5 also shows the ED Census generated from simulating the model under the baseline conditions.

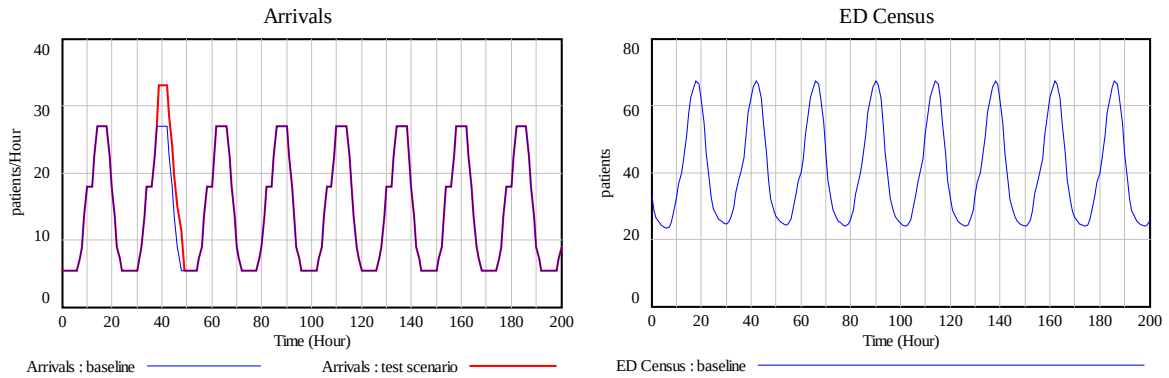


Figure 5. Left panel: Pattern of patient arrivals used as model inputs for the baseline and test scenarios. Right panel: Simulation results showing the total ED census in the baseline scenario.

To conduct simulation experiments with the model, we begin with the system in dynamic equilibrium as described and then introduce a change. For clarity of exposition, all of the simulations in this paper begin with the same initial conditions and then introduce at time=39 hours a one-time temporary increase in Arrivals that lasts for 10 hours after which Arrivals return to the original, baseline rate. The Arrivals graph in Figure 5 shows this surge of arrivals for one value (n = 6) of the temporary increase. The results of our first experiments, from introducing an increase of 5 patients per hour and 6 patients per hour, are shown in Figure 6. The first panel shows the Actual Wait Time for patients from the time an ED physician initiates the request for consult to the time the patient is transferred to a bed on the wards. The second panel shows the total ED Census, which is the sum of the stocks of Patients in ED plus those in Awaiting Consults, Awaiting Assign, and Awaiting Transfer. The results show the basic "physics" of the patients flows. At time 39, the increase in arrivals causes the ED census to begin to grow. Once the ED has stabilized and processed these patients, some are discharged and others are processed for admission. As the requests for admission begin to increase, the hospital beds become full. The Feasible Assign Rate drops well below the Admission Decisions and the stocks of patients Awaiting Assign and Awaiting Transfers grow. Arrivals slow

somewhat because of the diurnal pattern, bringing some relief in the congestion, but soon arrivals begin to grow again, causing the ED Census to grow as well. There are many patients still physically located in the ED, despite the fact that the ED physician and consulting specialist have already concurred to admit the patient and admitting orders have been written. Consequently, the Actual Wait Times grow. It takes quite some time for the effects of the surge in arrivals to dissipate, but they eventually do so, and over time the ED Census and Actual Wait Time returns to the original conditions. Recovery is slow, but the system has the resilience to eventually recover from the shock of additional arrivals. Figure 6 shows the results of another similar test when the magnitude of the temporary increase is 6 patients/hour. The results are qualitatively the same. These two simulations mimic the case of "access block" that has been described by other authors (Richardson 2002; Forster, Stiell et al. 2003; Liew and Kennedy 2003).

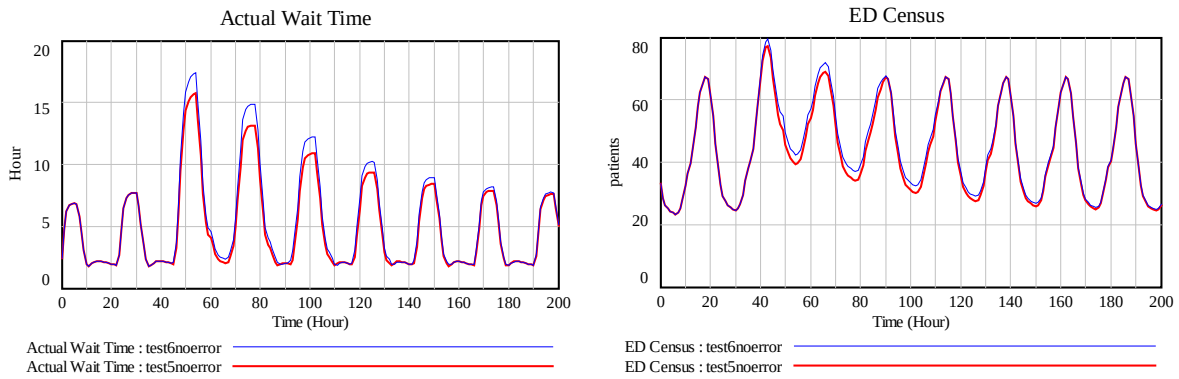


Figure 6. Response to a step increase in patient arrivals from time 10 to time 20 for a step height of 4 patients/ hour and a step height of 5 patients per hour.

Expanding the Model

The model in the previous section includes one important aspect of the physical constraints imposed on the ED by the fixed bed capacity of the hospital system within which it operates. In this section, we extend the model to encompass some behavioral effects of ED crowding. We include additional feedback loops in the extended model and then use it to conduct further

simulation analysis for the purpose of deepening our understanding of the dynamics of ED crowding.

The previous simulations show that constraints on bed availability cause patients to wait in the ED for extended periods. Under such conditions, the increased number of boarders in the ED results in a greater workload for the ED staff. We extend the model here to consider possible effects of the increased workload on patient management practices in the ED. There are many such possible effects, but here we explicitly represent just one. We consider the effects of workload pressure on the decision making associated with patient dispositions. Specifically, we assume that when workload gets significantly higher than the normal workload, some fraction of disposition decisions are different. Greater workload leads to a higher frequency of admissions decisions for patients that would not have been admitted under less stressful conditions - what we will call Admissions Due to Bias. These might occur because of mistakes made due to workload pressure, but they might also occur as cautious physicians facing demanding workload become more likely to lean towards choosing to admit a patient for whom the disposition decision is a rather close call - the Admission Bias increases. Greater workload can also lead to a higher frequency of discharge decisions for patients that would otherwise have been admitted - what we will call Discharges Due to Bias. To model the flows of patients with these dispositions due to bias, we adjust the stock and flow structure as shown in Figure 7. The stock of patients in the ED is now comprises a stock of Patients in ED Destined for Admission and a stock of Patients in ED Destined for Discharge. The physicians do not know a priori in which stock the patients belong, but for modeling purposes we track them separately. The figure also shows a stock of Potential Revisits that is increased by the flow of Discharges Due to Bias and decreased by the Revisit rate, as patients return to the ED through the flow of Pre-Admit Arrivals.

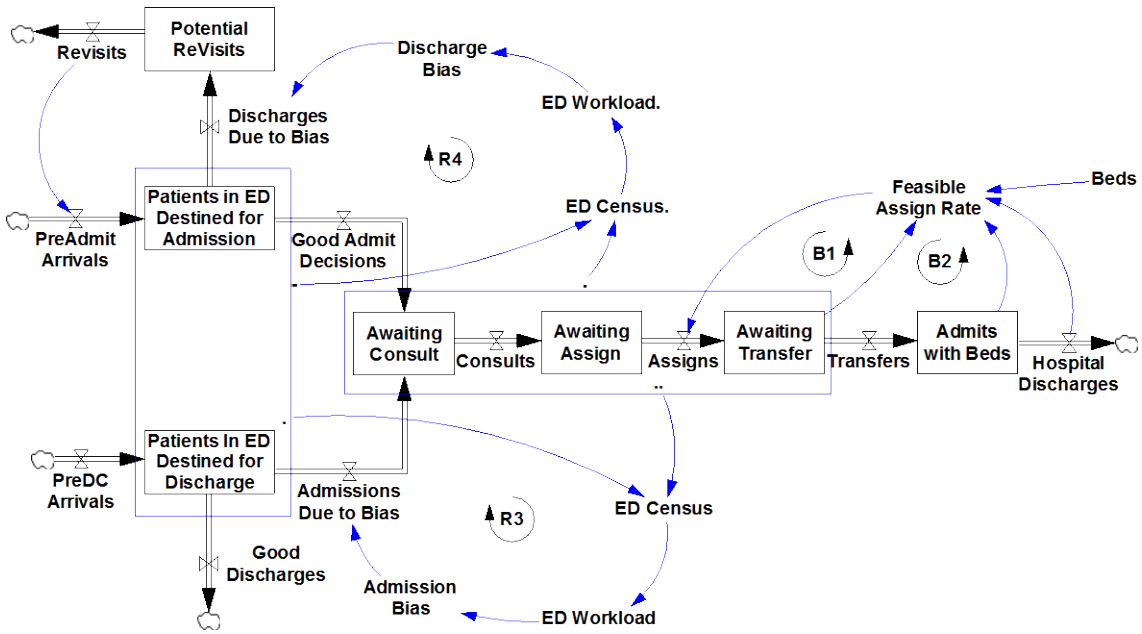


Figure 7. A model of patient flow in the ED showing constraints on bed availability and the effects of workload pressure on patient dispositions.

An important consequence of Admissions Due to Bias is that they increase the flow of patients generating demand for the admissions process of consult, assign, and transfer. When the bed constraints are binding, Admissions Due to Bias will cause an increase in the number of patients in the ED - and these patients still generate workload demands on ED personnel because the patients are still physically in the ED. The workload demand from a patient for whom the admission decision has already been made (i.e, a patient in the stock of Awaiting Consult, Assign, or Transfer) is considerably less than that from a patient who is still under active evaluation. Nevertheless, the former group of patients still draw on the ED resources. As shown in Figure 7, an increase in these stocks constitutes an increase in the ED Census, generating an increase in ED workload, which in turn cause the Admission Bias to climb, resulting in more Admissions Due to Bias and further increases in the stocks that form the ED Census. The feedback loop, labeled "R3," is a reinforcing feedback loop, because it acts to reinforce the direction of a change. Reinforcing loops move systems away from stability and are often implicated in dysfunctional dynamics.

To conduct our next simulation experiments, we need to specify the relationship between increased workload and the frequency of Admissions Due to Bias and Discharges Due to Bias. The effect of workload on disposition bias is modeled as an upward sloping nonlinear function of the actual workload compared to a threshold below which the bias is unaffected. For parsimony, we use the same effect functions for both admission and discharge biases (although the model allows us to parameterize these functions separately). Figure 8 shows how the Admission Bias depends on the variable Relative Workload, which is the current ED Workload compared to a threshold based on a multiple Normal Workload. Normal Workload is set to the peak workload experienced in the baseline scenario. The multiple of the Normal Workload is set to 1.05 in the following simulations. The tolerance of 5% additional workload above normal peaks before there is any effect on performance is a type of human capability that endows the ED with resilience to withstand a threat of increased demand. The Discharge Bias is model in exactly the same manner.

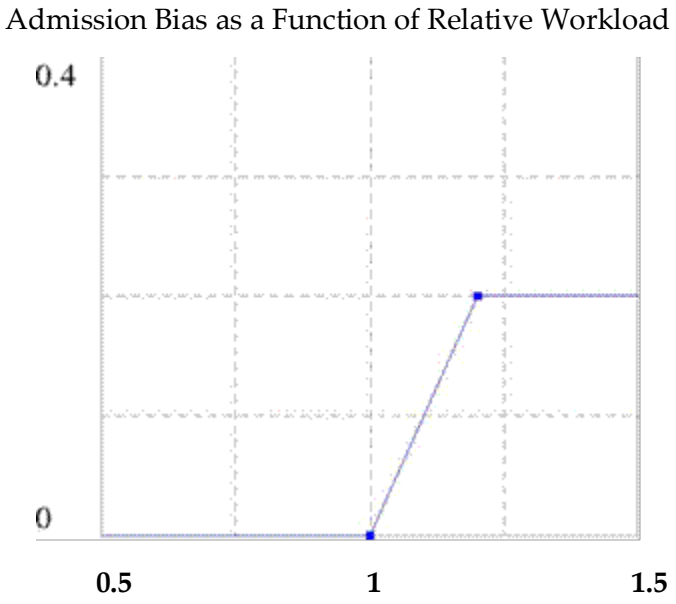


Figure 8. Admission Bias as a function of Relative Workload. Relative Workload is defined as the ratio of current ED Workload to the product of (1+Error Threshold) and the Normal Workload, which is defined as the peak workload in the baseline cycles. In the current model, the function for the Discharge Bias is identical

The blue line Figure 9 shows the response to a temporary increase in patient arrivals of five patients/hour. The upper left and upper right panels show the Actual Waiting Time and ED Census, as in the previous simulations. The lower left panel shows the Admission Bias, and the lower right panel shows the stock of Potential Revisits (which arise due to the Discharge Bias). The response of this stylized ED department, with physicians of finite capacity, appears from the salient metrics to be quite similar to the response shown in Figure 6 when there are no biases. The increase in arrivals soon leads to constrained bed availability, blocking access and causing ED Census to grow. Wait times grow as well. The system remains crowded for an extended period, taking six or seven daily cycles to fully recover as before, to the normal peak and trough census values. However, there are some weak signals that the system has been stressed if we examine the less salient Admission Bias and Potential Revisits. During the periods of peak census, workload is higher than the threshold for tolerating excess workload, so there are some Admissions Due to Bias and Discharges Due to Bias, as seen in the graphs of Admission Bias and Potential Revisits. Nevertheless, the system recovers, despite the challenge in the form of a burst of additional arrivals.

Next, we consider the response to a slightly larger surge in arrivals. The red line in Figure 9 shows the response to an increase of six patients/per hour. Although the most immediate response appears similar to that for the smaller surge, the ultimate behavior is quite different. The hospital fills quickly as before blocking access and causing a backup of patients boarding in the ED. As before, the additional workload demand from the growing ED Census leads to an increase in the disposition biases. But now, system performance deteriorates rapidly and continues to worsen even after the surge in arrivals is over. Although the Admission Bias begins to fall immediately once the surge in arrivals has subsided, the consequence of the Admissions Due to Bias during the period of peak excess demand remain in the system - literally as boarders in the ED - keeping workloads high. As the workload is still high enough to engender some Admissions Due to Bias among the ongoing arrivals of patients, there is continued inflow of patients in the Awaiting stocks greater than the feasible outflow to the hospital wards. The system here has crossed a critical threshold, or tipping point, and we see that ED Census and

wait times continue to grow. Growing census leads to more biased disposition decisions, which in turn increases the census, and the system behavior is swept into instability by this vicious cycle. The system is not able to recover from a shock of this magnitude, a shock which is only slightly larger than the shock shown in the blue line. In a real world system, at some point additional feedbacks would surely intervene, but this simulation highlights the potential vulnerability of the system. For a sufficiently large surge in arrivals, the system crosses a tipping point beyond which the reinforcing loop R3 in Figure 7 has come to dominate the system, and the system is permanently overwhelmed.

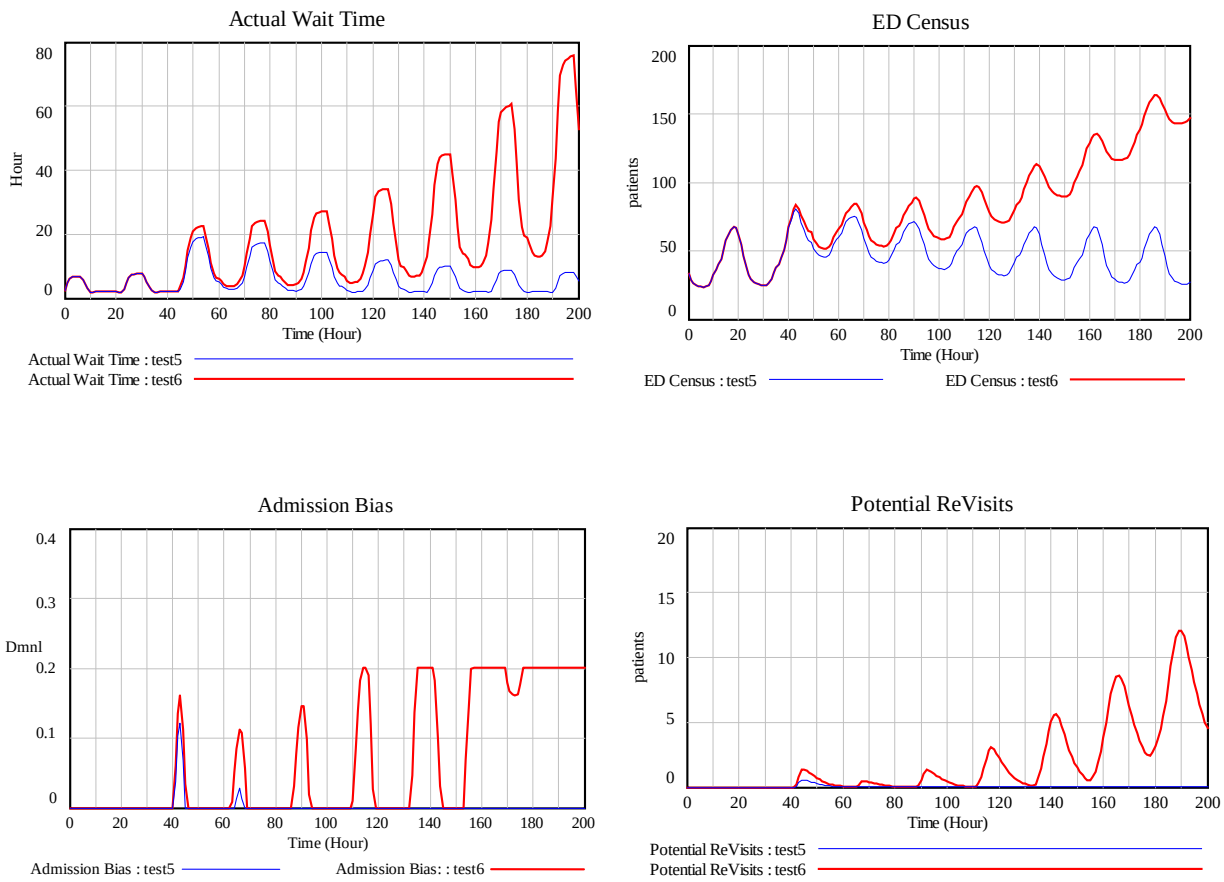


Figure 9. Simulations with the model shown in Figure 7. Response to a step increase in patient arrivals from time 39 to time 49 for a step height of 5 patients/ hour and a step height of 6 patients per hour.

To examine more closely the consequences of the biases in disposition decisions, we conduct the two simulations shown in Figure 10. The blue line shows the results when the only biases is the Discharge Bias decisions; that is, there are no Admissions Due to Bias. The red line shows the results when the only bias is the Admission Bias. The Admissions Due to Bias result in more patients in the ED, thus setting in motion the reinforcing loop R3. Discharges Due to Bias, in contrast, actually help the system by temporarily relieving some workload pressure. Although some fraction of these Discharge Due to Bias patients return to the ED, they leave during the period of extreme stress on the system. The model does not include adverse consequences on patient outcomes that no doubt arise from some Discharges Due to Bias, nor does it include an increase in the workload from a revisit patient that might be associated with the patient's worsening condition.

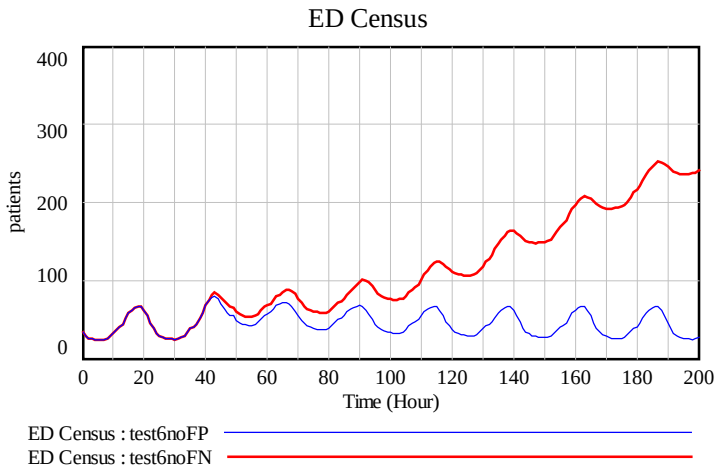


Figure 10. Simulations with the model shown in Figure 7. Response to a step increase in patient arrivals from time 39 to time 49 for a step height of 6 patients/ hour with no Admissions Due to Bias (blue) and no Discharges Due to Bias (red).

Next we turn our attention to some simulation experiments that examine the sensitivity of the system to characteristics of the physical environment and the behavioral responses. What if the ED physicians are more tolerant of the excess workload? To answer this question, we vary the parameter that sets the threshold workload above which biases begin. In the previous simulations, this threshold was 5% above normal peak workload. We test a small change in this threshold by setting it to 8% and show the results in Figure 11. For comparison, the blue line

shows the same simulation as the red line in Figure 9 - a response to a surge in arrivals of 6 patients/hour. The green line shows the response when the threshold for workload tolerance is 8%. The system is now able to respond effectively to the challenge from the surge in arrivals. With fewer Admissions Due to Bias, the ED avoids crossing the tipping point and they are able to recover once the surge in arrivals is over. These results highlight an important feature of the dynamics of this system. Human capability, such as the tolerance of the ED staff to excesses in workload, is sometimes able to overcome significant challenges to the smooth performance of the system. More insidiously, precisely because the human capability is able to do so, the signal that performance is threatened is muddled.

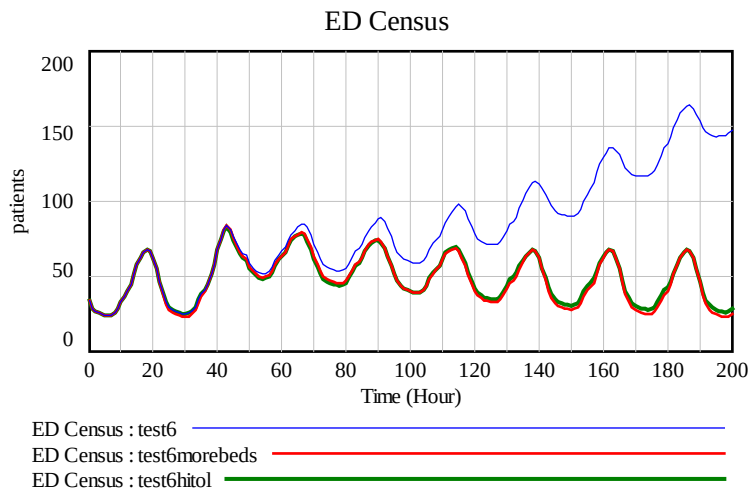


Figure 11. Simulations with the model shown in Figure 6. Response to a step increase in patient arrivals from time 39 to time 49 for a step height of 6 patients per hour. Blue: baseline. Green: Higher tolerances for workload stress. Red: Greater availability of beds.

Another possible improvement in this system is to free up more beds to be available for admissions from the ED. We conduct such a test in the model to see how the system behaves if there is easier access to beds by increasing the total number of beds. The grey line shows the system's response when the number of beds is two more than in the baseline scenario. The small increase in availability is enough to avoid the devastating overload, and the system is able to recover from the shock. This simulation demonstrates that, not surprisingly, changes in the physical environment (i.e., more beds) can make a system more resilient. The simulations in

Figure 11 highlight the important interaction between human capabilities and the physical environment. Both offer possible means for increasing resilience. A more physically robust workplace calls on less extreme human capability to achieve the requisite resilience to withstand a shock. Alternatively, a less robust physical setting requires more human capability to achieve the needed resilience.

The simulations in Figure 11 call attention to the interaction between patient management practices and the workplace setting in the hospital ED, highlighting that both dimensions have an important influence on patient flow dynamics and ED crowding. To further explore this critical interaction, we conduct a series of simulations in which we vary the size of the surge in arrivals (the input), the tolerance for excess workload (the human capability), and the number of beds (the physical setting). For several different combinations of bias threshold and bed availability, we conducted a number of simulations to determine the largest surge in the arrivals the system can withstand; that is, we identified the tipping points for each combination of parameters. The results are shown in Figure 12. Moving upward in this diagram represents increasing resilience - the ability to withstand and recover from a larger shock. For any given bias threshold (staying on any one line), greater availability of beds achieves greater resilience. Alternatively, for any given bed scenario (holding at one point on the horizontal access), increasing the bias threshold fosters greater resilience.

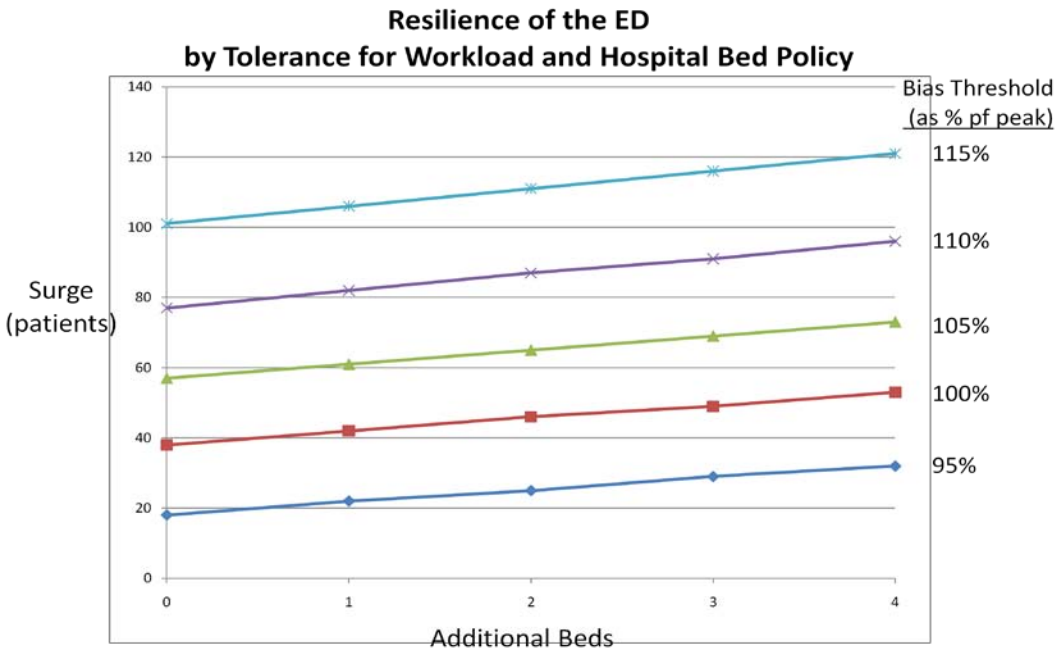


Figure 12. Results of experiments to identify the tipping points for various combinations of the Bias Threshold and Additional Beds. Results plot the influx (total number of additional patients over a 10 hour period) that pushes the system past the tipping point.

When hospital occupancy is high, the allocation of beds to ED patients is often difficult and occurs only after significant delays. We conducted a set of experiments to explore the effect of the timing of when an extra bed is made available. We use the same test scenario as before, a surge with an additional 6 patients per hour for 10 hours. Figure 13 shows the simulation results when no additional beds are allocated (blue line), which is the same as the blue line in Figure 11. The red line in Figure 13 shows the results when one additional bed is made available for ED patients 4 hours after the surge begins (t = 43 hours), and the green line shows the results when the additional bed is made available 8 hours after the surge begins (t=47 hours). The difference is outcomes is striking. When the allocation occurs 8 hours into the surge, the system does not recover from the surge. ED census levels are not as high as in the no extra bed scenario, but the census continues to grow long after the surge is over. The system has crossed the tipping point, and the additional bed allocated 8 hours after the surge begins is not adequate to resolve the situation.

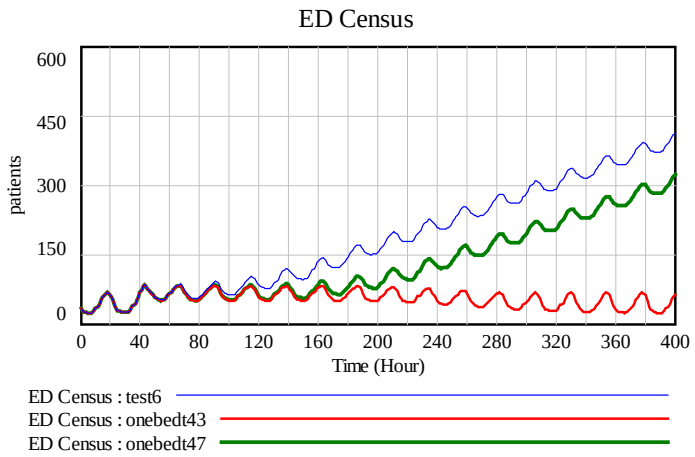


Figure 13. Results of experiments testing the effect of allocating extra beds. Response to a step increase in patient arrivals from time 39 to time 49 for a step height of 6 patients per hour. Blue: No extra beds. Green: One extra bed allocated 8 hours into surge. These two scenarios push the system past the tipping point. Red: One extra bed allocated 4 hours into surge. The system recovers.

Discussion

Hospital emergency departments are complex settings that bring together a mix of health care personnel in a dynamically changing environment with a changing mix of demands amidst significantly constrained resources, such as time and space. Most of the time, these emergency departments operate at remarkably excellent performance levels, even though most of the time it seems they are operating under extremely challenging conditions. This paper uses a system dynamics model to examine some aspects of patient flow dynamics in the ED. We show that beyond a certain point, the system loses its ability to recover from increases in demand in the form of excessive patient arrivals. The simulation results highlight that the vulnerability of this system is rooted in the critical interaction between physical constraints imposed by the environment (e.g., bed availability) and behavioral factors, such as the human capability of the ED staff to work at high performance levels under conditions of worsening workload pressure.

The simulation results mimic a quintessential feature of life in the ED. Staff in EDs face an increasingly challenging mismatch between demand for their services and their nominal capacity to provide such service. Yet, although there are some occasions of failure and some

signs of deteriorating performance, EDs across the country largely continue to avoid catastrophic collapse of their systems. Human capabilities (e.g., the physician's ability to continue to make proper dispositions in the face of adversity) compensate almost continuously for physical constraints and uncertainty. The simulation results show that increased ED resilience can come from relaxing bed constraints or from more human capability to cope. Importantly, there is a trade off of these two dimensions of bed constraints and workload tolerance. Improvement on one dimension can compensate for shortcomings in the other. In EDs where bed availability is constrained, staff that can tolerate extreme workload pressure without succumbing to disposition bias can enable the system to operate acceptably in response to greater shocks. However, more easy access to beds would enable the system to achieve the same levels of performance without the need to rely on the individuals who are more tolerant to workload excesses.

The concern arises because human capabilities are not infinite. When they get overloaded, system performance deteriorates rapidly. When operating near the tipping point these capabilities are the "buffer of last resort" that gives the system its resilience to recover. The human capability (of the ED staff in this example) to tolerate the extra workload masks the degree to which the bed constraint is threatening system performance, or at least reducing resilience.

Bibliography and References Cited

Adams, J. G. and M. H. Biros (2001). "The endangered safety net: establishing a measure of control." Acad Emerg Med **8**(11): 1013-1015.

Agrawal, S. (2007). "Emergency department crowding: an ethical perspective." Acad Emerg Med **14**(8): 750-751.

American College of Emergency Physicians (1990). "Hospital and emergency department overcrowding." Ann Emerg Med **19**(3): 336.

American College of Emergency Physicians (1990). "Measures to deal with emergency department overcrowding." Ann Emerg Med **19**(8): 944-945.

American College of Emergency Physicians (2006). "Crowding." Annals of Emergency Medicine **47**(6): 585.

Anderies, J. M., B. H. Walker, et al. (2006). "Fifteen weddings and a funeral: case studies and resilience-based management." Ecology and Society **11**(1): 21 - 32.

Asamoah, O. K., S. J. Weiss, et al. (2008). "A novel diversion protocol dramatically reduces diversion hours." Am J Emerg Med **26**(6): 670-675.

Asplin, B. R. and D. J. Magid (2007). "If you want to fix crowding, start by fixing your hospital." Annals of Emergency Medicine **49**(3): 273-274.

Asplin, B. R., D. J. Magid, et al. (2003). "A conceptual model of emergency department crowding." Annals of Emergency Medicine **42**(2): 173-180.

Baer, R. B., J. S. Pasternack, et al. (2001). "Recently discharged inpatients as a source of emergency department overcrowding." Acad Emerg Med **8**(11): 1091-1094.

Bagust, A., M. Place, et al. (1999). "Dynamics of bed use in accommodating emergency admissions: stochastic simulation model." British Medical Journal **319**(7203): 155-158.

Bayley, M. D., J. S. Schwartz, et al. (2005). "The financial burden of emergency department congestion and hospital crowding for chest pain patients awaiting admission." Annals of Emergency Medicine **45**(2): 110-117.

Burt, C. W. and L. F. McCaig. (2006, 27 September 2006). "Staffing, Capacity, and Ambulance Diversion in Emergency Departments: United States, 2003-04." Advance Data No. 376, 27 September 2006. Retrieved 16 November 2006, from http://www.cdc.gov/nchs/data/ad/ad376.pdf.

Cameron, P. A. (2006). "Hospital overcrowding: a threat to patient safety?" Med J Aust **184**(5): 203-204.

Cardin, S., M. Afilalo, et al. (2003). "Intervention to decrease emergency department crowding: Does it have an effect on return visits and hospital readmissions?" Annals of Emergency Medicine **41**(2): 173-185.

Committee on Pediatric Emergency Medicine (2004). "Overcrowding Crisis in Our Nation's Emergency Departments: Is Our Safety Net Unraveling." Pediatr Emerg Care **10**(2): 872-877.

Department of Health (2000). The NHS Plan: A Plan for Investment, a Plan for Reform. London, UK, London, UK: The Stationery Office.

Derlet, R., J. Richards, et al. (2001). "Frequent overcrowding in U.S. emergency departments." Acad Emerg Med **8**(2): 151-155.

Derlet, R. W. and J. R. Richards (2000). "Overcrowding in the nation's emergency departments: complex causes and disturbing effects." Annals of Emergency Medicine **35**(1): 63-68.

Diercks, D. B., M. T. Roe, et al. (2007). "Prolonged emergency department stays of non-ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events." Annals of Emergency Medicine **50**(5): 489-496.

Eckstein, M., S. M. Isaacs, et al. (2005). "Facilitating EMS turnaround intervals at hospitals in the face of receiving facility overcrowding." Prehosp Emerg Care **9**(3): 267-275.

Falvo, T., L. Grove, et al. (2007). "The opportunity loss of boarding admitted patients in the emergency department." Acad Emerg Med **14**(4): 332-337.

Falvo, T., L. Grove, et al. (2007). "The financial impact of ambulance diversions and patient elopements." Acad Emerg Med **14**(1): 58-62.

Fee, C., E. J. Weber, et al. (2007). "Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia." Annals of Emergency Medicine **50**(5): 501-509, 509 e501.

Forster, A. J., I. Stiell, et al. (2003). "The effect of hospital occupancy on emergency department length of stay and patient disposition." Acad Emerg Med **10**(2): 127-133.

Goldberg, C. (2000). Emergency crews worry as hospitals say, 'No vacancy'. New York Times. New York, NY**:** Section 1, pg 27.

Gordon, J. A., J. Billings, et al. (2001). "Safety net research in emergency medicine: proceedings of the Academic Emergency Medicine Consensus Conference on "The Unraveling Safety Net"." Acad Emerg Med **8**(11): 1024-1029.

Gould, S. J. (1989). "Punctuated equilibrium in fact and theory." Journal of Social and Biological Systems **12**(2-3): 117-136.

Gray, Z. A. and L. J. Baraff (2007). "The effect of emergency department crowding on time to parenteral antibiotics in admitted patients with serious bacterial infections." Annals of Emergency Medicine **x**(x): xx (in review).

Health Forum (2008). Hospital Statistics: 2008. Chicago, IL, American Hospital Association.

Hollander, J. E. and J. M. Pines (2007). "The emergency department crowding paradox: the longer you stay, the less care you get." Annals of Emergency Medicine **50**(5): 497-499.

Holling, C. S. (1973). "Resilience and Stability of Ecological Systems." Annual Review of Ecology and Systematics **4**(1): 1-23.

Hollnagel, E. and G. Sundström (2006). States of resilience. Resilience Engineering. E. Hollnagel, D. D. Woods and N. Levenson. Aldershot, UK, Ashgate**:** 339 - 346.

Homer, J. B. (1996). "Why We Iterate: Scientific Modeling in Theory and Practice." System Dynamics Review **12**(1): 1-19.

Homer, J. B. and G. B. Hirsch (2006). "System Dynamics Modeling for Public Health: Background and Opportunities." Am J Public Health **96**(3): 452-458.

Hoot, N. R. and D. Aronsky (2008). "Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions." Annals of Emergency Medicine.

Hoot, N. R., L. J. LeBlanc, et al. (2008). "Forecasting emergency department crowding: a discrete event simulation." Annals of Emergency Medicine **52**(2): 116-125.

Hwang, P. and J. D. Lichtenthal (2000). "Anatomy of Organizational Crises." Journal of Contingencies and Crisis Management **8**(3): 129-140.

Hwang, U. and J. Concato (2004). "Care in the emergency department: how crowded is overcrowded?" Acad Emerg Med **11**(10): 1097-1101.

Hwang, U., L. D. Richardson, et al. (2006). "The effect of emergency department crowding on the management of pain in older adults with hip fracture." Journal of the American Geriatrics Society **54**(2): 270-275.

Institute of Medicine (2006). Hospital-Based Emergency Care At the Breaking Point. T. N. A. Press. Washington, D.C., Institution of Medicine of the National Academies.

Jack, B. W., V. K. Chetty, et al. (2009). "A Reengineered Hospital Discharge Program to Decrease Rehospitalization: A Randomized Trial." Ann Intern Med **150**(3): 178-187.

JCAHO (2002). Delays in treatment. JCAHO Sentinel Event Alert.

Joint Commission on Accreditation of Healthcare Organizations. (2003). "Emergency department overcrowding standards." Retrieved 9 October 2003, from http://www.jcaho.org/accredited+organizations/hospitals/standards/draft+standards/er_fr_std.pdf.

Karpiel, M. (2004). "Improving emergency department flow. Eliminating ED inefficiencies reduces patient wait times." Healthcare Executive **19**(1): 40-41.

Kelen, G. D. and J. J. Scheulen (2007). "Commentary: Emergency department crowding as an ethical issue." Acad Emerg Med **14**(8): 751-754.

Kelen, G. D., J. J. Scheulen, et al. (2001). "Effect of an emergency department (ED) managed acute care unit on ED overcrowding and emergency medical services diversion." Acad Emerg Med **8**(11): 1095-1100.

Kellermann, A. L. (2000). "Déjà vu." Annals of Emergency Medicine **35**(1): 83-85.

Kellermann, A. L. (2006). "Crisis in the Emergency Department." N Engl J Med **355**(13): 1300-1303.

King, R. B., K. Shaw, et al. (2004). "ED overcrowding-meeting many needs." Pediatr Emerg Care **20**(10): 710-716.

Lane, D. C., C. Monefeldt, et al. (2000). "Looking in the Wrong Place for Healthcare Improvements: A System Dynamics Study of an Accident and Emergency Department." The Journal of the Operational Research Society **51**(5): 518-531.

Lattimer, V., S. Brailsford, et al. (2004). "Reviewing emergency care systems I: insights from system dynamics modelling." Emerg Med J **21**(6): 685-691.

Lesne, A. (2008). "Robustness: confronting lessons from physics and biology." Biol Rev Camb Philos Soc **(in press)**.

Levin, S., J. Han, et al. (2007). Stranded on emergency isle: Modeling competition for cardiac services using survival analysis. 2007 IEEE International Conference on Industrial Engineering and Engineering Management,. Singapore, IEEE**:** 1772-1776.

Levin, S. R., R. Dittus, et al. (2008). "Optimizing cardiology capacity to reduce emergency department boarding: a systems engineering approach." Am Heart J **156**(6): 1202-1209.

Liew, D. and M. P. Kennedy (2003). "Emergency department length of stay independently predicts excess inpatient length of stay." Med J Aust **179**(10): 524-526.

Litvak, E., P. I. Buerhaus, et al. (2005). "Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety." Joint Commission Journal on Quality and Patient Safety **31**(6): 330-338.

Litvak, E., M. C. Long, et al. (2001). "Emergency Department Diversion: Causes and Solutions." Acad Emerg Med **8**(11): 1108-1110.

Locker, T. E. and S. M. Mason (2005) "Analysis of the distribution of time that patients spend in emergency departments." British Medical Journal **330**, 1188 - 1189.

Lorch, S. A., A. M. Millman, et al. (2008). "Impact of admission-day crowding on the length of stay of pediatric hospitalizations." Pediatrics **121**(4): e718-730.

Moskop, J. C., D. P. Sklar, et al. (2008). "Emergency Department Crowding, Part 1: Concept, Causes, and Moral Consequences." Annals of Emergency Medicine.

Moskop, J. C., D. P. Sklar, et al. (2008). "Emergency Department Crowding, Part 2: Barriers to Reform and Strategies to Overcome Them." Annals of Emergency Medicine.

Nawar, E. W., R. W. Niska, et al. (2007). National Hospital Ambulatory Medical Care Survey: 2005 Emergency Department Summary. Advance Data from Vital and Health Statistics. Hyattsville, MD, National Center for Health Statistics.

Newton, M. F., C. C. Keirns, et al. (2008). "Uninsured adults presenting to US emergency departments: assumptions vs data." Journal of the American Medical Association **300**(16): 1914-1924.

Patel, P. B., R. W. Derlet, et al. (2006). "Ambulance diversion reduction: the Sacramento solution." Am J Emerg Med **24**(2): 206-213.

Pines, J. M. (2007). "Moving Closer to an Operational Definition for ED Crowding." Academic Emergency Medicine **14**(4): 382-383.

Pines, J. M. and J. E. Hollander (2007). "The Impact Of Emergency Department Crowding On Cardiac Outcomes In ED Patients With Potential Acute Coronary Syndromes." Annals of Emergency Medicine **50**(3, Supplement 1): S3.

Pines, J. M., A. R. Localio, et al. (2007). "The impact of emergency department crowding measures on time to antibiotics for patients with community-acquired pneumonia." Annals of Emergency Medicine **50**(5): 510-516.

Randers, J. (1980). Guidelines for Model Conceptualization. Elements of the System Dynamics Method. J. Randers. Cambridge, MA, Productivity Press**:** 117-139.

Rathlev, N. K., J. Chessare, et al. (2007). "Time series analysis of variables associated with daily mean emergency department length of stay." Annals of Emergency Medicine **49**(3): 265-271.

Reeder, T. J. and H. G. Garrison (2001). "When the safety net is unsafe: real-time assessment of the overcrowded emergency department." Acad Emerg Med **8**(11): 1070-1074.

Richardson, D. B. (2002). "The access-block effect: relationship between delay to reaching an inpatient bed and inpatient length of stay." Med J Aust **177**(9): 492-495.

Richardson, D. B. (2003). "Reducing patient time in the emergency department: most of the solutions lie beyond the emergency department." Med J Aust **179**(10): 516-517.

Richardson, D. B. (2006). "Increase in patient mortality at 10 days associated with emergency department overcrowding." Med J Aust **184**(5): 213-216.

Schneider, S., F. Zwemer, et al. (2001). "Rochester, New York: a decade of emergency department overcrowding." Acad Emerg Med **8**(11): 1044-1050.

Schull, M. J., A. Kiss, et al. (2007). "The Effect of Low-Complexity Patients on Emergency Department Waiting Times." Annals of Emergency Medicine **49**(3): 257-264.e251.

Schull, M. J., K. Lazier, et al. (2003). "Emergency department contributors to ambulance diversion: a quantitative analysis." Annals of Emergency Medicine **41**(4): 467-476.

Schull, M. J., L. J. Morrison, et al. (2003). "Emergency department overcrowding and ambulance transport delays for patients with chest pain." CMAJ **168**(3): 277-283.

Schull, M. J., J. P. Szalai, et al. (2001). "Emergency Department Overcrowding Following Systematic Hospital Restructuring: Trends at Twenty Hospitals over Ten Years." Acad Emerg Med **8**(11): 1037-1043.

Schull, M. J., M. Vermeulen, et al. (2004). "Emergency department crowding and thrombolysis delays in acute myocardial infarction." Annals of Emergency Medicine **44**(6): 577-585.

Shah, M. N., R. J. Fairbanks, et al. (2006). "Description and evaluation of a pilot physician-directed emergency medical services diversion control program." Acad Emerg Med **13**(1): 54-60.

Sprivulis, P., S. Grainger, et al. (2005). "Ambulance diversion is not associated with low acuity patients attending Perth metropolitan emergency departments." Emerg Med Australas **17**(1): 11-15.

Sprivulis, P. C., J. A. Da Silva, et al. (2006). "The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments." Med J Aust **184**(5): 208-212.

Sterman, J. D. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. Boston, Irwin McGraw-Hill.

Storrow, A. B., C. Zhou, et al. (2008). "Decreasing lab turnaround time improves emergency department throughput and decreases emergency medical services diversion: a simulation model." Acad Emerg Med **15**(11): 1130-1135.

US General Accounting Office (2003). Hospital Emergency Departments: Crowded Conditions Vary Among Hospitals and Communities. Washington, DC, US General Accounting Office**:** 71.

Viccellio, P. (2001). "Emergency department overcrowding: an action plan." Acad Emerg Med **8**(2): 185-187.

Viccellio, P. (2008). "Customer Satisfaction Versus Patient Safety: Have We Lost Our Way." Annals of Emergency Medicine **51**(1): 13-14.

Walker, B. and D. Salt (2006). Resilience Thinking: Sustaining Ecosystems and People in a Changing World. Washington, DC, Island Press.

Washington, D. L., C. D. Stevens, et al. (2002). "Next-day care for emergency department users with nonacute conditions. A randomized, controlled trial." Ann Intern Med **137**(9): 707-714.

Wears, R. L. and S. J. Perry (2006). Free fall - a case study of resilience, its degradation, and recovery, in an emergency department. 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France, Mines Paris Les Presses.

Weber, E., S. Mason, et al. (2011). "Emptying the corridors of shame: organizational lessons from England's 4-hour throughput target." Annals of Emergency Medicine **57**(2): (in press).

Weissman, J. S., J. M. Rothschild, et al. (2007). "Hospital workload and adverse events." Med Care **45**(5): 448-455.

Woods, D. D., J. Wreathall, et al. (2006). Stress-strain plots as a model of an organization's resilience. 2nd International Symposium on Resilience Engineering, Juan-les-Pins, France.

Worster, A., C. M. Fernandes, et al. (2006). "Identification of root causes for emergency diagnostic imaging delays at three Canadian hospitals." J Emerg Nurs **32**(4): 276-280.

Zwemer, F. L. (2000). "Emergency Department Overcrowding." Annals of Emergency Medicine **36**(3): 279.

Exploration de la dynamique de la performance résiliente

RESUME : Un thème récurrent dans les études sur la résilience est la nécessité de nouvelles méthodes de représenter les propriétés du système qui se concentrent sur les qualités dynamiques plutôt que statiques. L'objectif de cette thèse est de développer des modèles qui aperçu de soutien dans la dynamique du fonctionnement des systèmes résilients (et les gens entre eux) de gérer des situations instables. Il se concentre sur un enjeu spécifique, mais commune (surcharge), et sur les stratégies utilisées pour y faire face; en particulier, une stratégie spécifique, l'arrêt temporaire, afin de récupérer la marge de manoeuvre.

La thèse commence par une explication de l'étude de cas de surcharge de motivation sans exemple dans un service hospitalier d'urgence, conduisant à un effondrement sans précédent du système. Il analyse ensuite des cas similaires dans différentes configurations de prétendre qu'il ya des isomorphismes dans les stratégies et les adaptations aux différents niveaux et entre les domaines. Enfin, il développe un modèle de système dynamique d'un système général de travail en cas de surcharge, et l'utilise pour explorer les origines de la crise de surcharge, et l'utilité de la stratégie de l'arrêt temporaire de sa gestion. Elle montre qu'un indicateur avancé d'une crise imminente est l'impossibilité de recouvrer intégralement pendant les périodes normalement lent. Il montre également que l'arrêt est une stratégie risquée, et qu'il est facile pour les acteurs à apprendre les mauvaises leçons de leurs expériences. Ces résultats peuvent informer des moyens pratiques d'anticiper et d'atténuer les conséquences de la surcharge en milieu hospitalier et ailleurs.

Mots clés : Sécurité, la surcharge, l'adaptation, de résilience, l'arrêt

Exploring the Dynamics of Resilient Performance

ABSTRACT : A recurring theme in studies of resilience is the need for new methods of representing system properties that focus on dynamic rather than static qualities. The goal of this thesis is to develop models that support insight into the dynamics of how resilient systems (and the people in them) manage unstable situations. It focuses on a specific but common challenge (overload), and on the strategies used to cope with it; particularly, a specific strategy, temporary stopping, in order to recover margin for maneuver.

The thesis begins by an explication of the motivating case study of unexampled overload in a hospital emergency department, leading to an unprecedented system collapse. It then analyses similar cases from different settings to argue that there are isomorphisms in strategies and adaptations across levels and across domains. Finally, it develops a system dynamics model of a general work system under overload, and uses it to explore the origins of the overload crisis, and the utility of the temporary stopping strategy in managing it. It shows that a leading indicator of an impending crisis is the failure to recover fully during normally slow periods. It also shows that stopping is a potentially risky strategy, and that it is easy for actors to learn the wrong lessons from their experiences. These results can inform practical ways of anticipating and mitigating the consequences of overload in hospital settings and elsewhere.

Keywords : Safety, overload, adaptation, resilience, stopping

MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY