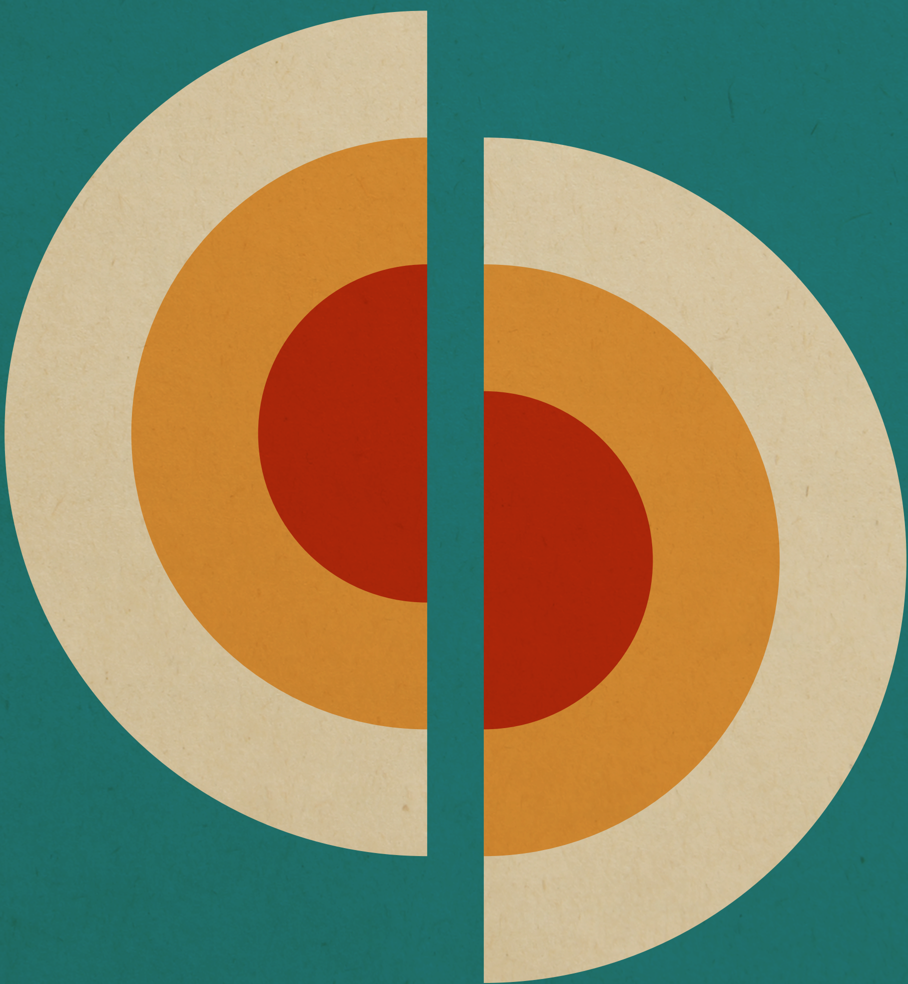
DIMITRIS NATHANAEL KOSTAS GKIKAS NICOLAS MARMARAS

Human Factors in Interactive Systems Design

a student's guide



KALLIPOS
open academic editions

National Development Programme 2021-2025

DIMITRIS NATHANAEL
KOSTAS GKIKAS
NICOLAS MARMARAS

Human Factors in Interactive Systems Design

a student's guide

KALLIPOS
open academic editions

Human Factors in Interactive Systems Design

Authors
Dimitris Nathanael
Kostas Gkikas
Nicolas Marmaras

Contributors/ Editors
Language Editor: Eleftheria Kanari
Graphics editor: Eleni Tsakmaki

Central Support team
Language Check: Eleftheria Kanari
Graphical Check: Elena-Natassa Kaitsa
Library Processing: Evdoxia Karli

Cover Design: Kostas Gkikas

KALLIPOS,
National Technical University of Athens
9 Iroon Polytechniou St., 15780 Zographou

www.kallipos.gr

ISBN: 978-618-228-173-4

Bibliographic Reference: Nathanael, D., Gkikas, K., & Marmaras, N. (2023). *Human Factors in Interactive Systems Design* [Postgraduate textbook]. Kallipos, Open Academic Editions. http://dx.doi.org/10.57713/kallipos-406

To our students

Table of contents

List of Figures with notes on copywrite

	Chapter 1
1.1	Photo by Alan Moore (CC BY-ND 2.0), https://www.flickr.com/photos/mrmoorey/150601428
1.2	Photo by João Lima, https://www.uiux.pt/2020/06/30/design-error-paul-fitts/
1.3	Own work, adapted from Regulation, F. A. 1964. Part 25: Airworthiness Standards
1.4-1.12	Own work
	Chapter 2
2.1-2.6	Own work
2.6	Adapted from Wickens C. D. (1992). Engineering psychology and human performance (2nd ed.). HarperCollins.
2.7	Adapted from Wisner, A. (1982). Physiologie du travail et ergonomie. Paris: Conservatoire National des Arts et Métiers.
2.8-2.13	Own work
2.14-2.19	Adapted photo from zoetnet, (CC BY 2.0), https://www.flickr.com/photos/zoetnet/4452175054/
	Chapter 3
3.1-3.9	Own work
3.10	Adapted photo from Nick Gray, (CC BY-SA 2.0), https://www.flickr.com/photos/nickgray/4334389323
3.11-3.19	Own work
3.20	Photo by Brandon Daniel, (CC BY-SA 2.0), https://www.flickr.com/photos/bdu/12211160175/
3.21-3.22	Own work
3.23	Photo by Jon Proctor (GNU Free Documentation License 1.2), https://www.wikiwand.com/en/Eastern_Air_Lines_Flight_401#Media/File:Eastern_Air_Lines_Lockheed_L-1011_Tristar_1_Proctor-1.jpg
3.24	Own work
3.25	Photo by the U.S. National Archives, Released to Public. https://catalog.archives.gov/
	Chapter 4
4.1	Photo by Thomas Quine (CC BY 2.0), https://www.flickr.com/photos/quinet/6091888784/
4.2-4.4	Own work
4.5	Photo Uncredited (CC BY-SA 4.0), https://commons.wikimedia.org/wiki/File:Wmalinowski_trobriand_isles_1918.jpg
4.6	Photo by Konstantinos Tsakalidis under permission
4.7-4.11	Own work

4.12	Photo by Thayne Tuason (CC BY-SA 4.0), https://commons.wikimedia.org/wiki/File:Walmart_Wenatchee.jpg
4.13	Figure by Tae Jung Yun et al. https://dl.acm.org/doi/10.1145/2110363.2110432
4.14	Own work
	Chapter 5
5.1-5.17	Own work
	Chapter 6
6.1-6.13	Own work
	Chapter 7
7.1-7.27	Own work
	Chapter 8
8.1-8.14	Own work
8.15	Figure by Transport for London, Public use
8.16	Figure by Esri UK Bureau (ODbL 1.0 license), https://hub.arcgis.com/maps/8a567ebac15748d39a747649a2e-86cf4/explore?location=51.514228%2C-0.108222%2C14.00
8.17	Figure by Transport for London, Public use, https://tfl.gov.uk/maps/track/tube
8.18-8.30	Own work
8.31	Photo by SJ Alice Bennett under permission
8.32-8.38	Own work
	Chapter 9
9.1	Photo by PH2 Milton R. Savage, Released to Public. https://commons.wikimedia.org/wiki/File:An_Agusta-Bell_212ASW_helicopter_assigned_to_the_Greek_frigate_HS_Limnos_%28F-451%29_returns_to_the_ship_after_carrying_a_boarding_party_to_a_merchant_vessel_for_an_inspection_during_Operation_Desert_Shield.JPEG
9.2-9.15	Own work

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
ATM	Automatic Teller Machine
CAD	Computer Aided Design
ECDIS	Electronic Chart Display and Information Systems
EEG	Electroencephalography
ESM	Experience Sampling Method
GPS	Global Positioning System
HAC	Human Action Cycle
HCI	Human Computer Interaction
HF	Human Factors
HF/E	Human Factors and Ergonomics
HMI	Human Machine Interaction
HTA	Hierarchical Task Analysis
ICD 10	International Classification of Diseases - Revision 10
ICT	Information and communications technology
ISO	International Organization for Standardization
IT	Information technology
LCD	Liquid-Crystal Display
LTM	Long-Term Memory
MD	Medical Doctors
MDF	Medical Call Form
MMAC	Maritime Medical Advice Center
NHTSA	National Highway Traffic Safety Administration
NTSB	National Transportation Safety Board
OCR	Optical Character Recognition
PC	Personal Computer
PIN	Personal Identification Number
SB	Service Blueprint
SCADA	Supervisory Control and Data Acquisition
SOP	Standard Operating Procedure
SRK	Skill-Rule-Knowledge
STD	State Transition Diagrams
STM	Short-Term Memory
STSS	Short-Term Sensory Store
TO	Telecommunications Operators
UCD	User Centered Design
UJM	User Journey Map
UX	User Experience

Preface

The present book is an introduction to Human Factors / Cognitive Ergonomics for interactive systems design, coupling applied theories of cognition with design methodology. The book came as a need to provide a comprehensive guide in user centered design destined primarily to students in information technology, engineering and design disciplines. The only prerequisite knowledge needed to follow the book is a basic understanding of engineering concepts; otherwise, the book introduces the necessary theoretical background in a progressive manner, thus a good understanding of previous chapters is a prerequisite on the ones that follow.

This work is the culmination of many years of experience in teaching Human Factors / Ergonomics (HF/E) together with relevant research and professional practice. Along with established and contemporary academic literature, this experience allowed the collection and development of a wide variety of examples and case studies of real-world systems from home appliances to large scale critical systems that the authors have found effective for transferring knowledge to students in a hands-on, experiential and dialectic manner.

The chapters are organized in three sections; chapters 1-3 provide a theoretical background in human cognition and action with relevant theoretical and applied models; chapters 4-8 roughly follow the user-centered design process of interactive systems, starting from the collection of user needs followed by requirements analysis, conceptual design, prototyping and iterative design / evaluation up to detailed interface design issues. Finally, chapter 9 is a complete case study presenting a specific user centered design process integrating all the previous phases. It should be noted though that the inherently linear structure of a book is not always compatible with the realities of a design process where iteration among phases is the norm. It is, thus, advised to follow the guide in a flexible, open manner and not consider it as a strict procedure. The readers can either use the book as an introductory guide to human cognition, a methodological companion for the user centered design process or as a collection of examples and case studies to enrich their repertoire.

The examples presented in the book refer mostly to everyday systems, and less to highly specialized and complex systems typically found in professional domains. This is a deliberate choice, as most readers have a good grasp and easy access to such systems if not direct experience. The downside of using everyday system examples is that having own experience from a system, tends to narrow one's ability to step into other users' shoes, and to recognize issues that they did not encounter themselves. In some cases, it may seem an overkill to apply such resource demanding methods to analyze and redesign trivial everyday products. Nevertheless, the system design process follows the same HF/E principles regardless of criticality or complexity. Therefore, the methods presented are applicable to a wider range of design efforts, from a ticket vending machine to a medical device or an aircraft cockpit. In addition, experience shows that these methods were a key ingredient in today's digital revolution and provided a head start to everyday systems/products that later dominated in their respective fields like the Google search engine, Apple's iPhone I or Amazon's e-shop.

Last but not least, in contrast to a typical handbook, the book is eclectic in the methods it presents. It does not strive for exhaustiveness, but instead, it opts for a demonstrative selection of methods and tools that have been found appropriate for educational purposes. Furthermore, it stresses the need for adapting methods to each particular project acknowledging the originality of each design endeavor.

Chapter 01
Introduction

Chapter 1: Introduction

Chapter Summary

The chapter provides the background of cognitive ergonomics going through the historical development of tool mediated human activity from direct physical manipulation of objects in the ancient world to executive control of modern fully automated systems. A series of examples is provided demonstrating the progressive distancing of human input from its effects in the environment through levels of tool mediation. Next, the scope and areas of application of Human Factors / Ergonomics is discussed, coupled with the evolving goals / criteria of success for ergonomic design and/or interventions.

Prerequisite knowledge

Basic knowledge of engineering concepts.

1.1 Interfaces and Human Factors/Ergonomics

Figure 1.1. The Cockpit of a B-17.

The Boeing B-17 (also known as the “Flying Fortress”) was one of the most successful warplanes against the Axis forces on World War II (Figure 1.1). Its robustness and airworthiness provided confidence to the pilots while they were flying through rain of bullets and shrapnel. But, it had a weak point that was far from the war zone, and more specifically at the end of each mission. In just two years, 457 of them crashed during their landing routine with no evidence of any mechanical malfunction or explanation from the surviving pilots. All the reports categorized the accidents as "pilot error". Just after the end of the war, the US Air Force called Paul Fitts (1912–1965), a psychologist at the Aero Medical Laboratory at Wright-Patterson Air Force Base, to investigate the reports. Along with his colleague Alfonse Chapanis (1917-2002), also psychologist and later industrial designer, did an in-depth analysis of the reports as well as field investigations on the actual B-17 cockpits. Their study revealed that pilots often mixed-up two identically shaped and proximal toggles, the one that commanded the landing gear and the other which controlled the wing flaps. That is to say, a tired pilot, under the stress of landing after combat, would accidentally retract the plane's wheels on landing thinking he was controlling the wing flaps to control speed, causing obvious catastrophe. (Figure 1.2). This repetitive "erroneous behavioral patterns", made the authors question the attribution of such instances as "human error" and introduce a new one: "designer error".

Figure 1.2. The two toggles, one for the landing gear and the other to control the wing flaps.

Unable to change the position of controls on existing aircrafts due to technical constraints, Chapanis came up with an ingenious yet simple solution: He created a system of skeuomorphic[1] knobs and levers that made it easy to distinguish the various controls of the plane. A pilot would, thus, visually map and feel the shape to the intended purpose minimizing the chance of confusion, even in the dark (Figure 1.3).

[1] Skeuomorphism is the concept of making elements of a design resemble their real-world counterparts.

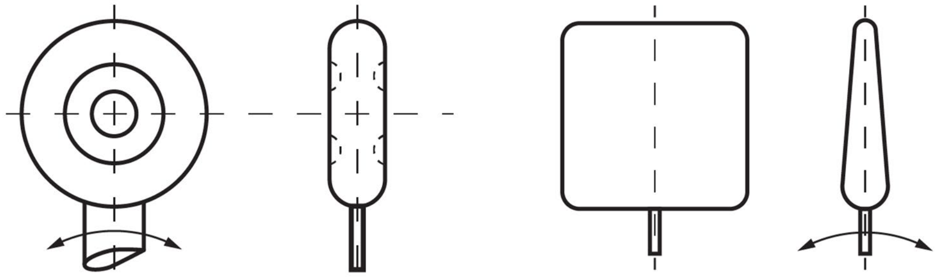


Figure 1.3. Distinctive shapes for landing gear and flap control knobs; not only do they look different but also feel different to the hand.

That design—known as shape coding—still governs not only landing gear and wing flaps in every airplane, but also every controller from vehicles and heavy machinery to controllers of video games. Paul Fitts and Alfonse Chapanis continued their career focusing primarily on aviation safety as well as other daily life user interfaces and they are considered today as two of the founding fathers of the Human Factors and Ergonomics (HF/E) discipline.

In 1974, another psychologist, John Voevodsky, equipped 343 taxicabs in San Francisco with a small, inexpensive apparatus that would eventually make driving much safer. The said apparatus was a third brake light, mounted in the base of rear windshields so that when drivers pressed their brakes, a triangle of light warned following drivers to slow down. Voevodsky randomly assigned taxicabs into two groups: one equipped with the third breaking light and one without (called the control group) and conducted a 10-month experiment recording the rear-end collisions for each group. His study showed that the taxicabs equipped with the extra brake light had suffered 60.6% fewer rear-end collisions than the control group. Additionally, drivers of taxis equipped with the third brake light that were struck in the rear by other vehicles were injured 61.1% less often than drivers of taxis without the light. The National Highway Traffic Safety Administration (NHTSA) repeated Voevodsky's experiment on a larger scale, and concluded that center high mounted stop lamps reduce accidents and injuries. As a result, in 1986, NHTSA began requiring all new cars to have a third brake light. Note that the addition of the third stop lamp not only increased breaking signal intensity of the car ahead (i.e., resulting in faster driver reaction) but often allowed for visibility of breaking activity of cars further ahead (Figure 1.4).

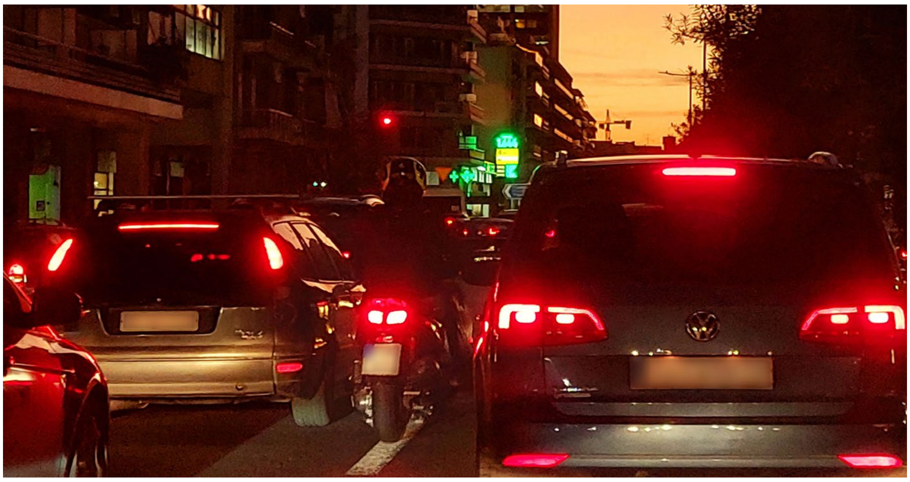


Figure 1.4. The third brake light, mounted in the rear windshield, became mandatory in cars after similar studies.

Both stories above highlight some key aspects of the Human Factors' approach, like the in-depth scientific knowledge of human abilities and limitations, extensive field work, thorough examination of the available data, and the proper experimentation of any new design idea. All in an attempt to create solid, failsafe interfaces that reduce errors and accidents, while promoting efficiency and pleasure of use.

In addition to high-risk systems as those mentioned above, the Human Factors' approach also found application in interactive systems for everyday use. While not likely to cause a major disaster, every day systems have a significant impact on the quality of people's lives as well as on market penetration and the cost of operation on a larger scale. Take, for example, e-government services or electronic transactions and e-commerce. As soon as such services became available to the wider public, many unexpected challenges arose. A study conducted back in 2000 by Claus M. Zimmermann and Robert S. Bridger on the use of Automatic Teller Machines (ATM) in banks, compared two existing ATM interfaces, and showed that lack of ergonomics input in one of the two, hereafter termed interface (a) was estimated conservatively at a loss of revenue of US$1.7 million due to poor task sequencing, and between US$2 million and US$4.5 million from user errors (mostly forgetting cards in the ATM). A fundamental difference between the two interfaces studied (a) & (b) was that (a) displayed the full range of functions, on the entrance menu including an exit option! To retrieve the card, the user had to command the machine by pressing the "no more transactions" option. On the other hand, interface (b) gave priority on withdrawals on entrance and placed all other functions in a separate menu, totally excluding the exit option. In interface (b) exit occurred without user input once a single withdrawal was completed, with the user's card being automatically dispensed just before the cash withdrawal. Card forgetting incidents were 96-100% higher when using interface (a) mainly because the routine task of card retrieval was sequenced after the user's primary goal was completed (i.e., receiving cash). In contrast to this, in interface (b) the user did not need to remember the routine card retrieval task since it was "a necessary step" before achieving his primary goal. Also, the researchers found that ATMs with interface (a) were overall 39% more time consuming (less efficient) to operate than (b) largely due to differences in dialogue design. As trivial as they may seem, such differences in dialogue design may have profound impact on system effectiveness and on user experience, not only in systems for everyday use but also in safety critical ones.

1.2 On Cognitive work

According to the Oxford dictionary, cognition, in general, refers to "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses". Many detailed scientific definitions of cognition can be found in the literature depending on the perspective and scope of each scientific discipline. In the present book, the following one best serves our purposes and will be endorsed.

"In a broad sense, the term cognition (or intelligence) implies at least the following mechanisms: the acquisition of environmental knowledge that can be improved through development and learning, the integration of information about past experiences, the development of strategies for adaptation of the individual to the environment and adaptation of the environment to the individual, automation of these strategies, prediction of future situations and strategies for responding to them "(Richelle and Dorz, 1976).

Stemming from the above, the term cognitive work is, therefore, used to describe facets of human activity related to (i) the collection of information from the environment, (ii) the processing of information in conjunction with knowledge already established in memory, including knowledge generation, (iii) the formation of intentions along with (iv) the neuromuscular control necessary to perform physical activities. Thus, it can be said that every human activity, even the most trivial, includes cognitive aspects, just as every cognitive work includes physical aspects (e.g., during speech). However, in some activities the cognitive aspects are more prominent than the physical ones and vice versa. It is what makes some tasks in everyday life to be characterized as "cognitive" or "mental[2] ", while others as "physical" or "manual". Nevertheless, any ergonomic study must consider both the mental and the physical aspects of work, as the two are closely intertwined and jointly define each activity.

[2] The term mental is broader than the term cognitive and refers to all processes related to the mind, including affective and emotional ones. However, no clear formal distinction can be made between the two and thus the terms will be used interchangeably in the text

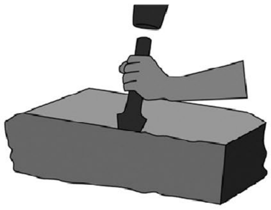
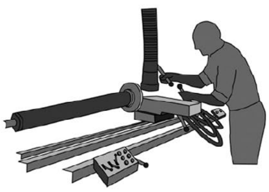
	Direct formation of tools by hand and by other materials
	Formation of artifacts using simple tools
	Formation of artifacts using simple man-powered machines: • simple machine settings • direct handling of the machine • easy machine repairing
	Manufacturing of objects with automatic power machinery: • manipulation through controls (the physical fatigue is reduced) • monitoring of proper functioning through indicators • decisions for corrective actions
	Production processes remotely supervised with extensive use of automatic control: • scheduling of production • control through preprogrammed procedures • detection and management of unforeseen situations by automatic control systems

Figure 1.5. The historical evolution of the relation between humans and work. Gradual distancing from the material process has made the work less strenuous physically but has increased the cognitive load of the worker.

In the modern era, there is an ever increasing use of automation and information technology in most work settings, and also in people's daily lives. As a result, the demands of cognitive work are becoming increasingly important. Thus, for example, industrial workers are increasingly moving from being handlers of material processes and product assemblers, towards becoming programmers of complex machines and supervisors of their proper operation (Figure 1.5). They are called to intervene when "something goes wrong", diagnose the causes of the anomaly and restore the proper functioning of the technological apparatus. These tasks require continuous reception and processing of information, prediction of the evolution of complex dynamic phenomena, decision making, planning and scheduling of actions etc.

This gradual evolution in work requirements triggered the development of the field of Cognitive Ergonomics (often also termed Cognitive Engineering). Cognitive Ergonomics deals with the mental components of work combining knowledge and methods mainly from Systems Theory, Cognitive Psychology, Psycho-linguistics, Social/ Cognitive Anthropology and Ethnography. Typical applications of Cognitive Ergonomics concern:

- user-centered design of man-machine interfaces (including human-computer interfaces)
- design of systems to support complex cognitive tasks (e.g., supervisory control, decision making, diagnosis, design),
- the design / redesign of work content and methods aiming to reduce mental workload,
- the study of human error and the improvement of human reliability,
- the development training programs based on cognitive analysis.

1.3 The Evolution of Interface Design

All tools crafted or fabricated by humans in our long-civilized history may be seen as consisting of two main parts: the working part and the handling part. The working part is shaped to act in the environment, while the handling part is designed so that someone can manipulate the tool. For example, a knife generally consists of a sharp and tough side that is used to cut the respective material and a rounded palm friendly side so as to grab it (Figure 1.6). Even the first stone tools used in the prehistoric era had these distinct parts. Most recent tools like robotic arms or automobiles also have the necessary parts to “do the actual work”, whether this is industrial welding or transporting people, and also some parts that are dedicated to their operation by humans, like a robot control pad or the wheel and the gear lever. All parts that fall on the second group are considered the interface of the tool or machine.

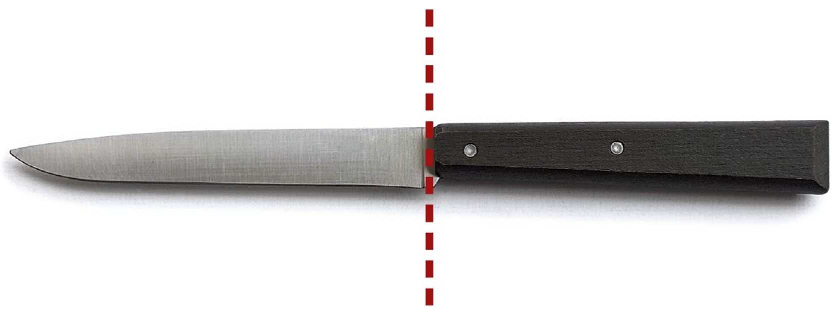


Figure 1.6. The line that separates the handling part from the working part.

Through developments in technology and societal needs, the design of tools tends to evolve over the years so as to become more task efficient. As a case in point, take a look at early versions of a weighing scale. The interface of this device –the balance weighing scale– consists of the two suspended metal plates, on which the material in question and the calibrated weights need to be placed, but also the horizontality (leveling) of the lever over the fulcrum to allow "reading" the balance. These are the parts that someone will use to measure an unknown mass. In subsequent versions the interface evolved even though the operating principle remained the same. The metal plates for materials became larger to facilitate handling of liquids and grain products (like flour, rice etc.), while the suspending chains were removed for easier fill. Also, the leveling of the lever was later replaced by a needle for better precision in reading (Figure1.7).

Figure 1.7. Two early versions of a weighing scale.

Further evolution of the weighing scale focused primarily on the need for even greater precision, but also on efficiency of operation, by replacing the external calibrated weights with springs inside the main body of the scale. At this evolutionary stage, the interface consisted of a single metal bowl and a magnified scale screen. Note that progressively all the moving parts along with the operating principle started retracting inside the body of the device (Figure 1.8).

Figure 1.8. Two more sophisticated mechanical scales.

Most contemporary weighing scales are electronic with their operating principle (usually compression load cells) fully hidden from view. In fact, the only part of the device visible to the user is the interface still consisting of a container and a reading element (now in the form of digital screen) (Figure 1.9). This evolution made the weighing task faster and more precise, while the device is more inexpensive to make, lighter and easier to store. In a sense, one could say that this version solved all previous inconveniences in a way that the weighing task became as easy and efficient as it can get.

Figure 1.9. A contemporary weighing scale.

However, in this optimization lies the biggest challenge that we face today in most advanced technological systems. The effort to offload the user from the often-complex underlying operating principles, although desirable, also has a major drawback. It hinders the user's ability to diagnose and solve possible malfunctions. In the above digital scale for instance, the user is unable to check whether the device functions correctly (e.g., the scale may correctly measure the value "8" but the screen may display the digit "2" due to a defective LCD), or more so how to solve such a problem even if correctly diagnosed. Note that erroneous readings on a spring based mechanical weighing scale can never be too far from the correct reading due to the hing scale can never be too far from the correct reading due to the physical coupling of the device elements, whereas in a digital device, erroneous readings may be totally random.

Computer programming languages present a similar pattern in their evolution. From a human factors' perspective, programming languages are regarded as intangible interfaces, their function being to mediate (allow communication) between a human and a digital algorithmic machine. The need to program a machine that works based on electrical signals by humans, who use mostly symbolic notions to communicate, led to a series of interfaces (i.e., programming languages) with successive levels of "translation". The process has gone from Machine Code to low level programming (assembly languages) and from there to high level programming (C, Fortran, Pascal) and so on (Figure 1.10). Each new level allowed for faster programming, fewer coding errors and steeper learning curves. However, the downside was that each successive level became opaquer and added more programming constrains to the users, than its predecessor. This aspect of higher-level languages made it more difficult to diagnose and rectify errors when programming.

Nevertheless, the scalability of programming environments makes it possible to write code with a high-level language for efficiency, but may allow the programmer to revert to a lower-level language, whenever in need for more detail and coding freedom.

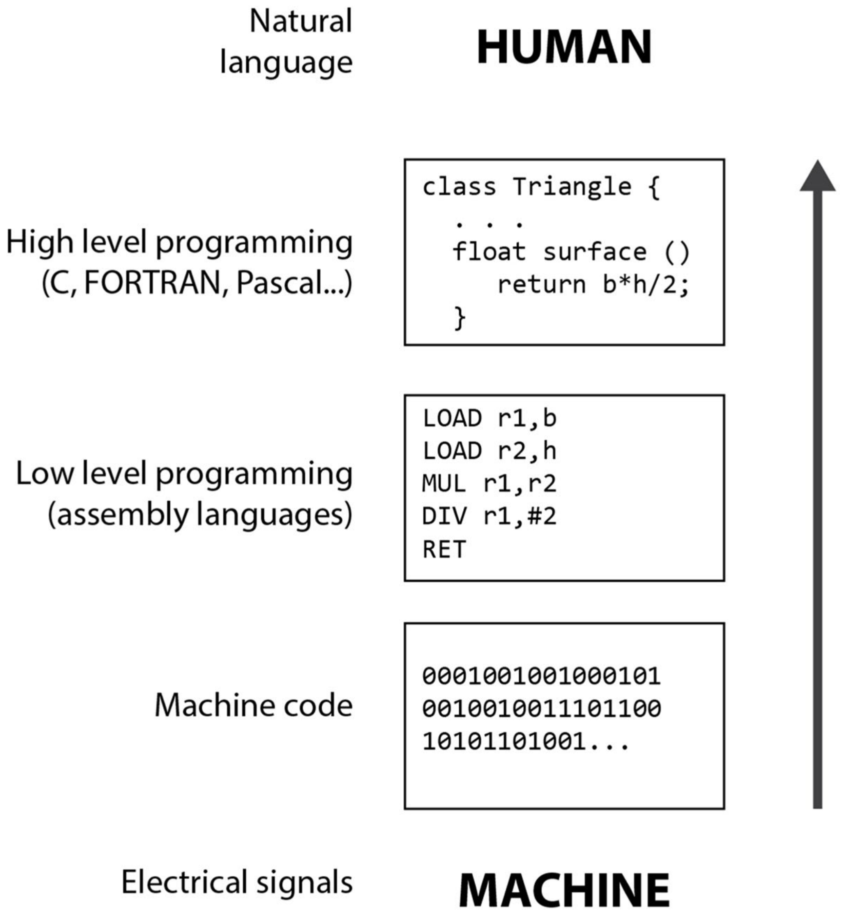


Figure 1.10. The several levels of programming.

Even generic examples, as the above, make evident that the more hidden the functioning of a device or tool at the interface level is, the more the user has difficulty in overseeing it.

This phenomenon is accentuated in the “interfaces” of large-scale safety critical systems like airplane cockpits or industrial process control rooms. In such systems, the opacity and complexity of inner system functioning at the level of the human interface have resulted in severe accidents over the years including aviation, petrochemical and nuclear disasters. So, are we more in danger today than we were in the past? The answer is “no” because contemporary devices and systems are far more technically reliable than in the past, and cause undoubtedly less accidents pro rata. Modern systems do not fail as often as older ones, but in the rare cases that they do, it may become almost impossible to diagnose and intervene in due time. In the next chapters we will shed more light as to why this happens and how we can mitigate such complexity – visibility tradeoffs through good design practices.

1.4 Examples of current advances in the field

Current advances in Information and communications technology (ICT) have led to new ways of interaction with technological systems, and new perspectives in the collaboration between humans and machines. Industrial robots cooperating with human workers, autonomous cars sharing the same streets with human drivers, and surgeons conducting operations remotely, are few of the current challenges. On many occasions, in order to understand the complex design issues that emerge from these changes we must rethink and analyze ordinary but often unacknowledged aspects of human behavior.

A case in point is the advent of autonomous cars that are expected to share the same streets with human drivers and pedestrians in the near future. How will these machines interact with humans and vice versa? Until recently the tacit communication schemes among drivers to coordinate their trajectories had not been thoroughly studied because such need did not arise. Now these tacit social communication schemes must be analyzed and formally modelled in order to be embedded in autonomous car algorithms and sensors along with ways to transmit and receive intentions, prompts and acknowledgments.

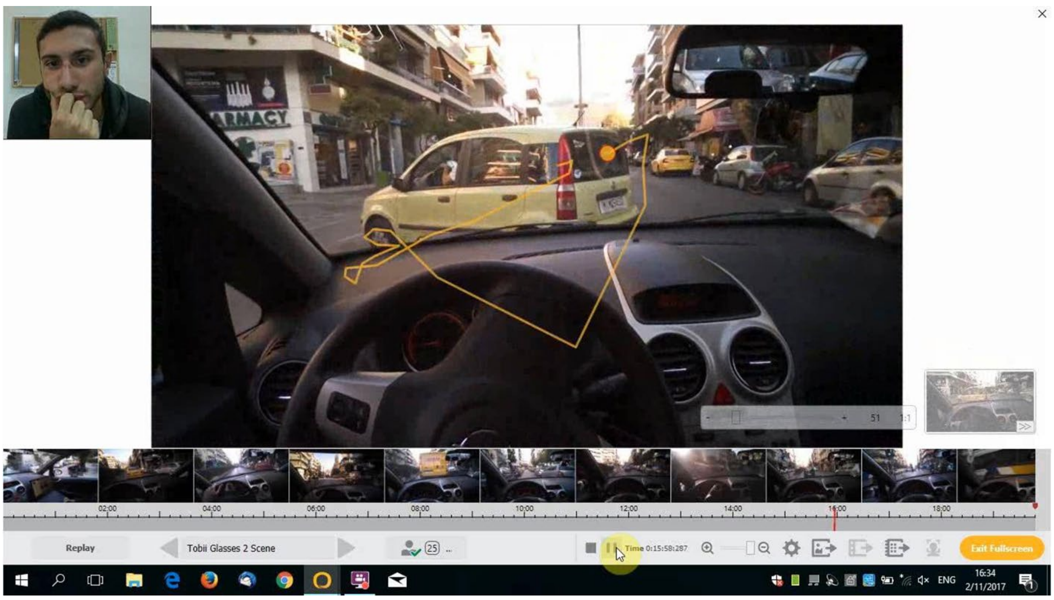


Figure 1.11. A driver commenting while watching his eye-gaze video recording.

In an eye-tracking study conducted by our lab on experienced drivers in urban environment (Figure 1.11), three respective levels of behavioral expressiveness from the part of the pedestrians were recognized for solving ambiguities on the road. Specifically, it was found that drivers were first looking for pedestrians' body cues, then for gazes towards the vehicle and ultimately eye contact with or without gesture signals, so as to make decisions concerning pedestrian intentions. Depending on the level of signal received, the drivers responded with more or less confidence. The above observations permitted the development of a stratified model with different states of driver pedestrian interaction based on the level of mutual awareness (Figure 1.12).

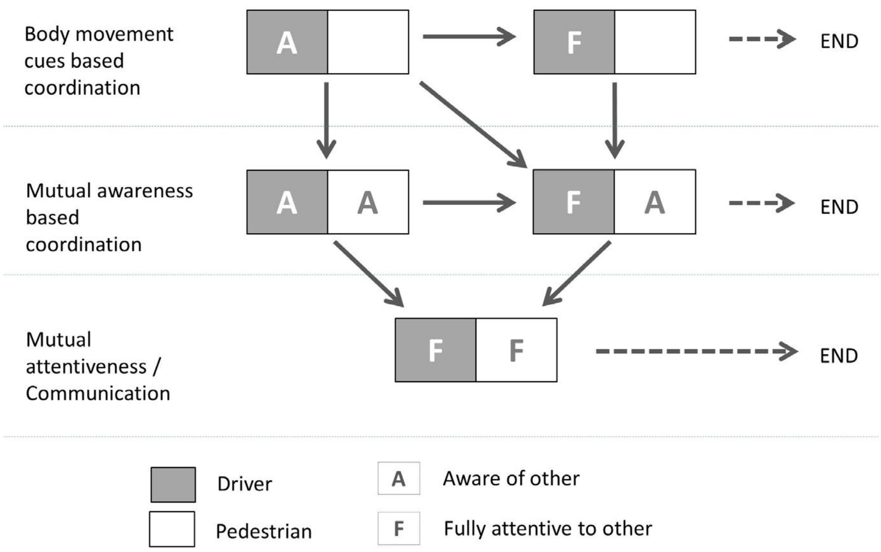


Figure 1.12. Model of possible states of mutual attentiveness between driver and pedestrian from the driver's point of view.

As it is seen in Figure 1.12, at the start of an interaction, a driver becomes aware of a pedestrian; this may be followed by the driver fully attending (i.e., gazing) to the pedestrian. In both these two possible states of drivers' attentiveness, an interaction may be effectively accomplished solely through physical movement co-ordination, e.g., relying on pedestrian body cues (i.e., body movement cues-based co-ordination). However, in more demanding cases of time/space resource sharing, mutual attentiveness becomes necessary. Drivers typically use pedestrians' eye gaze towards their vehicle as a cue to confirm pedestrians' readiness to co-ordinate and share the same resources; this tactic being often sufficient for completing an interaction, saving perceptual resources (i.e., mutual awareness-based co-ordination). Only when this mutual co-ordination leads to vagueness of intent or misunderstanding, both road users are forced to devote their full attention (through eye-contact, and hand gestures) and to communicate explicitly to each other their future intended actions (i.e., mutual attentiveness/communication) (Nathanael et al., 2019).

Recognizing such cues and acting accordingly is crucial for autonomous cars as we cannot expect from pedestrians or drivers to change their already established road behaviors shortly. Moreover, in cases where there is a need for explicit communication, an autonomous car should be capable of receiving but also emitting proper signals that could be easily understood from human road users regardless of cultural background or environmental conditions. Such challenges are confronted nowadays by the Human Factors discipline internationally, as the technology advances rapidly and the need for implementation of such solutions is imperative.

In the next two chapters, we will introduce a brief theoretical background in human cognition and action along with relevant applied models and concepts proper to Human Factors / Ergonomics. A good grasp of these theories, concepts and models is deemed prerequisite before proceeding to subsequent chapters which focus mainly on methodological aspects of Interactive Systems Design. The readers are advised to revisit these theoretical chapters throughout the study of the book, whenever reference is made to them, to refresh their memory and strengthen their understanding.

References

Chapanis, A. (1999) *The Chapanis chronicles: 50 years of Human Factors research, education, and design*. Aegean Pub Co.

Kuang, C. & Fabricant, R. (2019) *User Friendly: How the hidden rules of design are changing the way we live, work, and play*. MCD.

Lima, J. (2020) "Designer Error" – The first brick from the User-Friendly world. uiux.pt. https://www.uiux.pt/2020/06/30/design-error-paul-fitts/

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2019). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20* (pp. 389-397). Springer International Publishing.

Regulation, F. A. (1964). Part 25: Airworthiness Standards: Transport Category airplanes, https://www.ecfr.gov/current/title-14/chapter-I/subchapter-C/part-25/subpart-D/subject-group-ECFR9bfdfe36b332e4a/section-25.781

Chapter 02: Elements of cognition

Chapter 02: Elements of cognition

Chapter Summary

In this chapter the basic theories of human cognition are briefly introduced. These theories are considered necessary for the presentation of applied models and methods that follow later in the book. Specifically, first, the fundamental cognitive processes of short-term and long-term memory are introduced, followed by sensing/ perception and elements of cognitive processing. Integrating the above, next, the information processing model is presented along with a discussion on the limits of its application. Focus is then turned to semantics, introducing elements on the theory of signs, followed by contemporary ecological and embodied approaches to studying cognition. It should be noted that research on human cognition is still quite active and thus, many of the concepts introduced are in perpetual questioning, and as a result, often not in line with each-other. In any case, an effort was made to combine well-established theories with newer knowledge in a simplified but – as far as possible – coherent manner that was judged useful for design purposes.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of the previous chapter.

2.1 Human memory

A fundamental function of cognition is memory. Through the integration of experience in our mind, we are able to remember past experiences, recognize similar situations and develop new more effective action strategies each time, depending on our goals. Through the process of memory, each integrated experience acts as a stepping stone helping us adapt and develop.

2.1.1 The short-term memory / working memory

Memorizing ability is nevertheless constrained by various limits that need to be considered when designing systems requiring a certain level of cognitive effort to operate. To get a grasp of such limits, consider the following game.

In a live audience of at least 20 participants, show the following sequence of digits for 20 seconds, and ask the participants to memorize as many as they can without using any recording aid:

4 7 3 1 9 4 0 5 1 6 2 8 2 4 1 8 3 7 2 4 6 5 2 9

After the 20 seconds period, hide the sequence and ask the participants to write down as many digits as they can in the correct order, starting from left. You then reveal the sequence again and ask the participants to check how many digits of the sequence, starting from left, they got correct. Having run this test for more than 10 years with our students, we have noticed that the results follow a normal distribution with a mean value of 8 and standard deviation of 4. These numbers are very close to the theory that sets the limits of the so-called short-term memory in 7 (±2) independent chunks of information, proposed by psychologist George Miller (1956) in an influential paper titled "The Magical Number Seven, Plus or Minus Two".

Although there is still debate as of the exactness or even the relevance of this limit in terms of a fixed number of items, short-term memory can be generally defined as the ability of holding a small amount of information in the mind, in a readily available state, for a short period of time. For example, short-term memory is used in remembering a phone number that has just been recited. The natural duration of short-term memory (without rehearsal) is believed to be in the order of seconds. For example, memory decay can occur in less than 20 seconds for verbally delivered navigation information (Loftus et al., 1979). Information in short-term memory is also highly susceptible to interference. Any new information that enters short-term memory will quickly displace old information.

But how is it that some of our students surpass the theoretical limit of 7 (+-2) by a considerable margin? First, being between 22 and 24 years old, our students stand out of the general population in terms of cognitive abilities. But even young age does not justify scores over 11 that we often witness in the classroom. The secret behind such performance lies mostly in strategy. Top scoring students tend to memorize the sequence as a row of two- or three-digit numbers (47 31 94 ... or 473 194 ...) which seems to provide significant advantage over the memorization of single digit numbers. In fact, nonexceptional performance in all these years have been achieved without using such strategies. Since our short-term memory has a limited amount of space for storing information, as we populate it, less room remains for additional content. Therefore, the number "47" seems to take up less space in our memory than the numbers "4" and "7" do separately. The key point here is that the short-term memory limit refers to "chunks of information" and not mere units. This implies that a two-digit number might be considered a single chunk of information, somewhat more complicated than a single-digit chunk, but nevertheless more efficient in extending our memorizing ability. From our classroom experiments, we have seen that the two-digit chunking is the optimal solution for most people between the overload of single digits sequence and the complexity of the three-digit chunks. This memory technique, known as "chunking," is indeed commonly employed when memorizing phone numbers or passwords. Other observed strategies include the opportunistic matching of a sub-sequence with familiar patterns (e.g., part of one's social security number) or the rehearsed verbalization of the sequence as a poem. In fact, very often, before proceeding with brute memorizing, we tend to spend some time trying to identify any tricks that may alleviate this strenuous task. This is achieved by a process called central executive that manipulates two hypothetical information maintenance subsystems, the phonological loop and the visuospatial sketch pad. According to Baddeley and Hitch (1974) the three subsystems above form the construct of working memory.[1]

As a case in point, you may continue the experiment with the below sequences:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37

2 4 8 16 32 64 128 256 512 1024 2048 4096 8192

You will notice that there is, in fact, no real need to memorize those. With just four seconds of screening, most literate people can replicate the whole sequence on a piece of paper. This is because either these sequences are already processed in Long-Term Memory (LTM), or they readily derive from simple mathematical operations in working memory.

Finally, you can try with the following sequence:

3 3 5 4 4 3 5 5 4 3 6 6 8 8 7

At first, it seems as a series of unrelated digits, so you will most probably decide that it is not worth spending time searching for a potential pattern. In reality, those digits do follow a straightforward algorithm, so that, if someone has adequate time to spend and is quite observant, they might be able to recognize it. Each digit is, in fact, the total count of letters of the integers in English (one, two, three, four, five, six, seven, eight, etc.). Knowing the rule you can replicate the sequence at any given time, meaning that you only have to memorize the rule and not the sequence.

[1] Theoretically wise the concepts of short-term memory and working memory are not ideφοοτνοτεntical. Short-term memory is more of a simple store, while working memory allows it to be manipulated. In this sense, short-term memory can be considered as a part of the working memory. Nevertheless, there is still debate concerning their differences that go beyond the scope of the present textbook.

While it is empirically known from experimental studies that the capacity of STM is finite, the memory games presented above are meant to demonstrate that people will inherently try augmenting its functional capacity, in various ways. Indeed, in naturally occurring situations, people will devise tricks to optimize information chunking, develop cognitive artifacts (e.g., rules) or even inscribe information in the physical environment. What is even more crucial is that the functional capacity of the STM cannot be fully isolated from the operation of LTM.

Consider a well-known example from the literature; Chase and Simon (1973) studied the ability to memorize the positions of the pieces on a chessboard. The pieces were not randomly placed but represented actual games in progress. The participants in the experiment were of three levels: Master level, A-level, beginner. The results of the experiment showed that a Master player could correctly recall the position of about 16/24 pieces after seeing the chessboard for 5 seconds, while an A level player could correctly remember the position of only 8/24 pieces and a novice player only 4/24 pieces after looking the chessboard for the same duration. In contrast, when in the same games in progress, some pieces were randomly interchanged on the chessboard, Master players' recall score was often poorer than that of a beginner. The reasonable hypothesis for the interpretation of these results is that Master chess players can reconstruct from their LTM typical meaningful configurations of a game and therefore, retention in STM of a limited number of pieces is enough for them to generate through their LTM a quite plausible overall picture of the chessboard. When piece configuration is no longer meaningful (i.e., random piece replacement) Master players straggle more, because their effort was less on blind memorizing in STM than on trying to figure out meaningful configurations based on their LTM.

2.1.2 The long-term memory

The idea of distinguishing human memory into short-term and long-term goes back to the 19th century. A classical model of memory developed in the 1960s assumed that all memories pass from a short-term to a long-term store after a small period of time. There is a common belief, largely influenced by the advent of information technology, that human memory resembles to an information storage unit akin to digital memory devices. In recent years, however, this long-lasting belief and associated model has been largely abolished. We now know that humans do not exactly store and retrieve information in the way a digital computer does. On the contrary, it is more productive to think of human memory not so much as storage and retrieval of information but better as a process; a process of reconstructing experiences and formerly conceived ideas that have left loose traces in the mind.

A bicycle, for example, is a ubiquitous object in most peoples' lives and readily recognizable as such. But can the image of a bicycle be as readily retrieved from people's memory? Consider this simple test; try sketching a bicycle right out of memory without looking at one. You will notice that, unless you are an experienced biker, it is a rather challenging task, even though you have certainly seen numerous bicycles in your life and its form is rather simple. In addition, if you are asked to repeat this task some days or weeks later, your sketch will most likely not be the same. Each time you try to reconstruct the image of a bicycle the result will be slightly different (Figure 2.1). What is also interesting is that not all bicycle parts are difficult to recall. Almost everyone correctly reproduces the two wheels, one in front of the other, the saddle, pedals and handlebar on top of the front wheel. The trickiest part lies in the frame. However, if you are called to choose between various sketches of bicycle frames where only one is correct, you will directly recognize it.

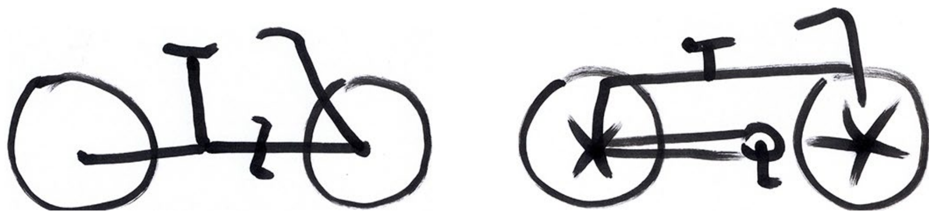


Figure 2.1. Examples of bicycle drawings sketched from memory by adults.

The frame is clearly an integral and visually inescapable part of any bicycle, but it is of minimal practical interest when one is riding it. The significant parts for riding include the saddle, the handlebar, the pedals, and the wheels. These are indeed the ones that are typically correctly sketched in the right place. The above example stresses the fact that our attentive resources and memory are highly selective. We tend to focus on what we experience as important and leave the rest as a fuzzy image, which we will try to reconstruct if needed, filling it with assumptions and guesswork.

A similar test that we run each year with our Greek students is the challenge to draw the map of Crete, the largest and most distinctly shaped island of Greece (Figure 2.2). We first separate the students into the ones originating from Crete, the ones having recently visited the island, and the ones that have never set foot on it. We, then, ask one of each group to sketch the map of the island as detailed as possible, strictly from memory. In most cases, it is obvious which sketch belongs to the respective group. Sketches from native Cretans are typically much more detailed than the others'. However, they also tend to be selectively distorted by exaggerating on the size and detail of the coastline of their birth region compared to the rest of the island. The same goes to a lesser degree for those recently having visited a certain part of the island, but in diminished detail, while those who have never been there give a very rough sketch of an amorphous beam shape. It goes without saying that all students readily recognize the actual map when shown to them. Recollection of information is, therefore, much more demanding cognitively than recognition. The simple task of remembering a familiar shape is essentially a productive process mingling bits and parts from diverse cognitive resources such as visual traits, episodic events, and abstract concepts. Recollection resembles, therefore, more to a cognitive act rather than a request of retrieval of stored information.

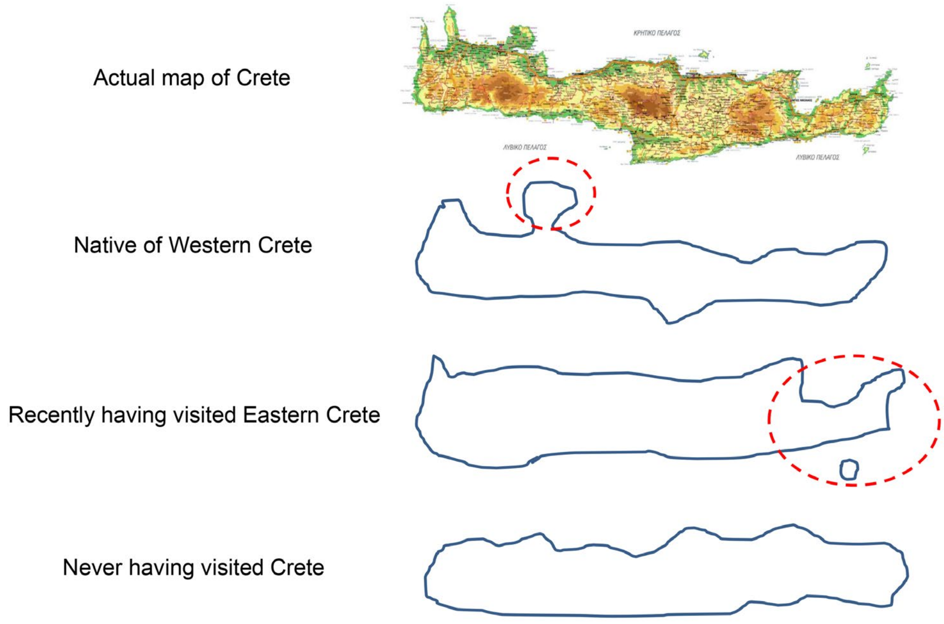


Figure 2.2. The actual shape of the island of Crete (top) and typical student sketches from memory depending on their familiarity with the island (three bottom sketches).

Elizabeth F. Loftus (born 1944), an American cognitive psychologist and expert on human memory, took these ideas further and conducted research on the malleability of human memory. Loftus is best known for her work on the misinformation effect (1989) and eyewitness memory (1979), and the creation and nature of false memories (2013). Overall, she has shown that the human memory is dynamically changing information acquired, based on a person's beliefs, values and psychological condition, and that it can also be manipulated and altered. Her findings have changed the role of witnesses in legal systems and gave new perspectives on cognitive science.

The way in which information and/or knowledge are organized in LTM, is not fully known. Various hypotheses have been proposed and partially confirmed in the experimental laboratory; however, it is still not possible to propose a robust general model. Some of the strongest assumptions about how information is embedded in memory are: (i) temporal or local bundling (information that arrives in time sequence or from the same point in the environment is memorized in the same bundle), (ii) semiotic hierarchical structures (an example of organizing concepts in memory is shown in Figure 2.3), (iii) mental schemata, which are structures gradually created developmentally and in which, all knowledge or information related to experiences reside (Piaget & Inhelder, 1968).

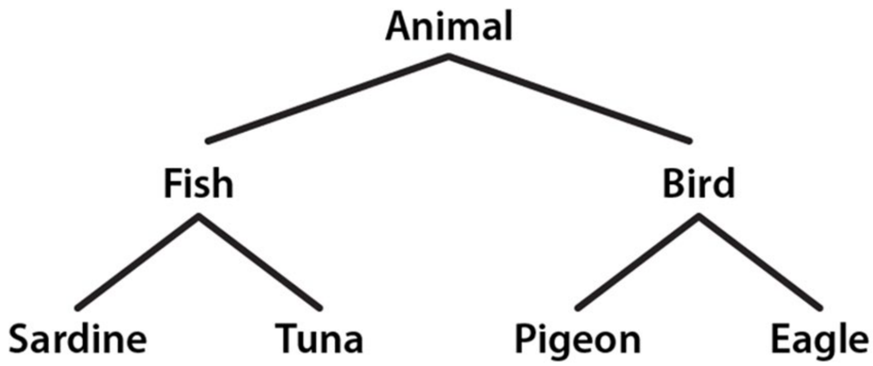


Figure 2.3. A semiotic hierarchical structure.

Long-term memory is commonly distinguished according to how it manifests itself as (i) declarative memory (explicit), i.e., memory that can be articulated through words, and (ii) procedural memory (implicit) i.e., non-articulable memory, such as the skill of tying shoes or driving a car (Stillings et al., 1995). This distinction is also supported by neuroscience findings that locate these two manifestations of memory reside in different parts of the brain (Blakemore, 1988).

Declarative memory entails knowledge that can be communicated to others in words or in an indicative proposition (Ten Berge & Van Hezewijk, 1999). In that sense, declarative memory is conscious by definition, as it presupposes intentional recollection of factual information, previous experiences, and concepts. This type of memory seems to be controlled by the hippocampus which is a complex brain structure embedded deep into the temporal lobe. Declarative memory can be further divided into two broad types: "episodic memory" and "semantic memory". Episodic memory entails recollections of experiences such as the day we started college and our last birthday party. This type of recall is our interpretation of an episode or event that occurred to us. On the other hand, semantic memory entails factual knowledge, such as recalling the names of European countries or the form and parts of a car, without clear connection to specific life episodes.

Procedural memory, on the other hand, entails knowledge that is manifested in the performance of some familiar task. It is better understood as know-how or skill. So, for instance, the knowledge of how to use a technological device, such as a car, a coffee maker, a mobile phone or a computer application, is part of procedural knowledge. One does not need to be able to verbally articulate their procedural knowledge for it to count as knowledge. Its existence is only demonstrated by correctly performing an action or exercising a skill. Procedural knowledge is often also called "tacit knowledge". Cerebellum, basal ganglia and motor cortex are involved in procedural memory, which is overall supervised by the cerebral cortex. Hippocampus, which is essential for actively controlling the explicit memory, is not needed during activation of implicit memory. However, for the implicit memory to form, explicit memory has to form first and train the cerebellum and other centers. Thus, for the formation of new implicit memory the presence of an intact hippocampus is a necessity. (Dharani, 2015). The distinction between these different kinds of memory can explain why amnesiac patients, who have lost their ability for conscious recollection of recent events, not only do they retain all their previously acquired skills and basic knowledge of the world around them but are also able to acquire new skills through practice.

Procedural knowledge does not seem to fully reside in the brain; parts of this “knowledge” can be said to extend all around the body. For example, in riding a bicycle, all parts of the body “learn” how to maintain balance while pedaling, constantly performing micro movements, and shifting rhythmically the body weight from one side to the other. A simple, yet demanding, choreograph, inscribed at many levels, from the upper layers of consciousness to the last neuronal loops and muscles of our hands and toes. What is more, this knowledge cannot be transferred from one person to another through words, sketches or even mimicry. Due to its embodied nature, it can only be learned through own experience, making it a tacit skill embedded in the body memory of each rider[2]. It is evident from the above that declarative knowledge (i.e., description) of a process or skill is not sufficient for its successful conduct. One can have detailed declarative knowledge on how to ride a bicycle without being able to actually put it to effect.

2.2 Sensation and perception

The sensory system is the part of the nervous system dedicated to sensing and processing sensory information. It consists of sensory neurons (including the sensory receptor cells), neural pathways, and parts of the brain involved in sensory perception. In short, senses are transducers from the physical world to the realm of the mind, where we interpret the information, creating our perception of the world around us.

While debate exists among contemporary neurologists as to the specific number of senses, due to differing definitions of what constitutes a sense, five ‘traditional’ human senses have been universally accepted from ancient times (see classifications of both Gautama Buddha and Aristotle): touch, taste, smell, sight, and hearing. Other senses that have been well-accepted in most mammals, including humans, include nociception, equilibrioception, kinesthesia, and thermoception. Furthermore, some

[2] For a more detailed account on the role of the body in cognition, see the section on embodiment in this chapter below.

nonhuman animals have been shown to possess alternate senses, including magnetoception and electroreception (Hofle et al., 2010). All sensory receptors have some mechanisms in common, such as detection, amplification, discrimination, and adaptation.

The triggering of sensation stems from the response of a specific receptor to a physical stimulus. The receptors which react to a stimulus and initiate the process of sensation are commonly characterized in five distinct categories: (i) chemoreceptors (detecting certain chemical stimuli) as in taste and smell, (ii) photoreceptors (detecting a portion of the light spectrum and convert it to membrane potential) as in vision, (iii) mechanoreceptors (responding to mechanical forces such as pressure) as in touch and hearing, (iv) thermoreceptors (responding to varying temperatures). All receptors receive distinct physical stimuli and transduce the signal into an electrical action potential. This action potential, then, travels along afferent neurons to specific brain regions where it is processed and interpreted.

Sense receptors retain the stimuli received momentarily in an isomorphic way. Thus, it is considered that each sensory system incorporates a Short-Term Sensory Store (STSS) of its stimuli. Experimental data show that information is imprinted on the STSS, even if the individual's attention is focused elsewhere. The STSS is generally limited, in terms of capacity and duration, and many experiments have been performed over the years to quantify these limits. However, there is no unanimity regarding the numerical limits of STSS, and these limits seem to have a very wide range of values. Some indicative values for the capacity of STSS are 7 – 17 random letters that do not form words for the visual STSS and 4.4 – 6.2 random letters for the acoustic STSS. Concerning the duration of information that stays in the STSS, the various experiments show a range of 70 – 1000 msec for visual stimuli and 900 – 3500 msec for auditory stimuli (Card et al., 1986). Note that these stimuli are still isomorphic to their external causes, an unprocessed visual image or sound.

At a second stage, selected stimuli are categorized and signified through the process of perception. Perception is the organization, identification, and primary interpretation of sensory information. It enables individuals to form a coherent model of the world around them, even though the sensory information is typically incomplete and variable.

Perception is activated by the received stimuli but is also heavily moderated by the recipient's memory, expectations and attention. Specifically, the stimuli characteristics that will be perceived as useful information depend mainly on the receiver's preoccupation (i.e., intentions and expectations) and on the knowledge and experiences that are imprinted on LTM, and allow an individual to match stimuli to them. For example, the red signal collected by a driver's visual system and imprinted on the STSS is forwarded to the STM, so the driver becomes aware of it, identifies it as a traffic signal through prior experience.

Perception is, in fact, a quite complex process with interconnected modules that influence each other (e.g., perception of taste is heavily influenced by smell). However, to the individual, it seems mostly effortless because it largely happens without conscious awareness.

2.3 Cognitive Processing

Cognitive processing refers to various types of cognitive operations carried out in the creation and manipulation of mental representations of information. These include general functions such as reasoning, problem solving, decision making, leading up to imagination. Human cognitive processing is extremely complex, multifaceted, and variable to be covered by a general descriptive model. Nevertheless, it presents some invariant properties that differentiate it from purely rational, algorithmic information processing. In the following subsections, we briefly explore some of these properties mainly to stress the differences of cognitive processing from analytical reasoning.

2.3.1 Hicks law

Figure 2.4. The number of choices affects decision time and choice quality in a non-trivial manner.

The design of decision support systems for complex cognitive tasks is one of the core interests of the Human Factors & Ergonomics discipline. To better design such systems, it is critical to understand some particularities of the human decision making process in simple tasks. After all, being able to make the right decisions can be considered a critical aspect of intelligence. Consider the case where you walk into a coffee shop to take your morning coffee beverage and you face a list of 9 different choices you have never seen before. Suppose it takes 20 seconds to make a decision. If the choices were 18, as in Figure 2.4, would you do double the time to decide? What about 36 choices, would that take you 80 seconds, less or more? When deciding between two alternatives (A, B) we generally compare some aspects of them, and when deciding between three objects (A, B, C) we compare the same aspects between A-B, B-C and A-C. But do we really make all the possible comparisons when the number of objects increases? That would mean that the possible comparisons and thus, the time to decide increase exponentially. But does it really happen so?

In 1952, William Edmund Hick, a British psychologist and Ray Hyman, an American psychologist, set out to examine the relationship between the number of stimuli present, and an individual's reaction time to any given stimulus. What they found was that the choice reaction time increases along with the logarithm of the number of alternatives. The law is usually expressed by the formula RT = a + b log2 (n), where a and b are constants representing the intercept and slope of the function, respectively, and n is the number of alternatives (see Figure 2.5). The above formula is widely known as Hick's Law (or the Hick-Hyman Law). As someone would expect the more stimuli to choose from, the longer it takes the user to decide on which one to interact with. But, what is more important is that those factors do not have a linear correlation. That means that, if the available choices increase, let's say, from 20 to 40, the time to decide and react will increase but certainly not be doubled. In fact, after a certain number of choices the reaction time will remain almost the same. So, does that mean that the increase of choices cost proportionally less than the benefits we have over that increase? Certainly not. Hick's law deals only with the decision time and not the quality of this decision. The increase in potential options overwhelms our ability to make comparisons between all options, so we inevitably make shortcuts to unload our cognitive work. In front of 40 options when buying a coffee, we will quickly reduce them to a number of choices –probably three to six– with which we feel comfortable to compare and decide. Doing so will save us time and cognitive effort, but our decision will be compromised, having avoided dealing with most of the available choices, as if they didn't exist at all.

Therefore, offering many options to users (even if there is no time pressure) will not only increase decision time up to a certain point, but most importantly, will passively compel users to make less-informed choices.

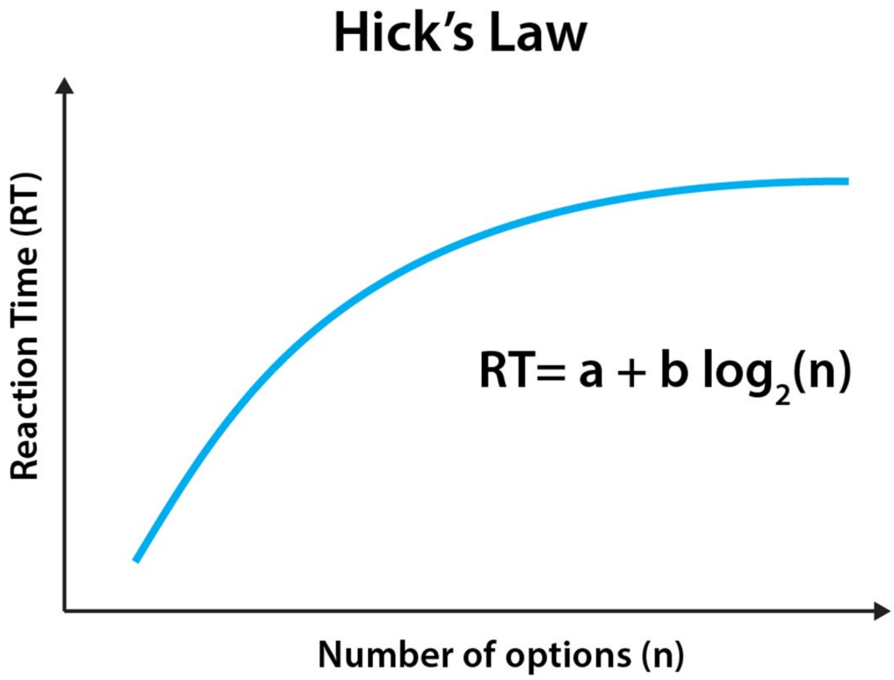


Figure 2.5. Hicks Law. Increasing the number of choices will increase the decision time logarithmically (a and b are constants, |a| representing the time not involved with decision making and |b| the cognitive process time per option; |n| is the number of alternatives).

2.3.2 Cognitive Biases

Suppose that we have chosen a rule that some sequences of numbers obey, and some do not. You are asked to guess what the rule is. To help you, we are giving a sequence that obeys the rule:

2 - 4 - 8

Before answering what the rule is, you can test your hypothesis by proposing some new three number sequences to test if these sequences obey our rule or not. Most respondents would propose sequences such as the following: 16 – 32 – 64 / 1 – 2 – 4 / 3 – 6 – 12 / 3 – 6 – 18, trying to verify either a rule of doubling each number, or multiplying the two numbers to get the third and so on. In any case, if the reply on whether the new sequence obeys the rule was positive, they would make a new try to verify further their guessing, and then they would reveal the rule that they had in mind with much confidence. Truth is that all the above sequences obey the hidden rule, but none of the proposed rules is correct, simply because the rule is much simpler: each number must be larger than the previous one. What is interesting in this example is not whether someone figured out the correct rule or not, but in the way most people have tried to test their hypothesis. In most cases, people make an assumption of what the rule is and try to test their assumption by selecting sequences aiming to confirm it and not to disprove it. Choosing one over the other method gives no real advantage; either way, when we have to test a hypothesis, we can either try to verify it or refute it. Remarkably, 77% of people who have played this game on a post of the NY Times, have guessed the answer without first hearing a single “no”. A mere 9% heard at least three “no’s” even though there is no penalty or cost for being told “no”, save the small disappointment that every human being feels when hearing “no.” This disappointment is a version of what psychologists call confirmation bias. Not only are people more likely to believe information that fits their pre-existing beliefs, but they are also more likely to go looking for such information. This experiment is a version of one that the English psychologist Peter Cathcart Wason used in a seminal 1960 paper on confirmation bias.

The following cognitive biases are considered some of the most common ones that relate to our ability to process and interpret information in the world around us and affect the decisions and judgments that we make.

- **Confirmation bias**
  the tendency to process information by looking for, or interpreting, information that is consistent with one’s existing beliefs or values.
- **Anchoring or focalism**
  The tendency to rely too heavily, or ”anchor”, on one trait or piece of information, when making decisions (usually the first piece of information acquired on that subject).
- **Attentional bias**
  The tendency of perception to be affected by selective factors in their attention (e.g., by recurring thoughts).
- **Automation bias**
  The tendency to depend excessively on automated systems which can lead to erroneous automated information overriding correct decisions.
- **Framing effect**
  Drawing different conclusions from the same information, depending on how that information is presented.
- **Observer-expectancy effect**
  When a researcher expects a given result, and therefore, unconsciously manipulates an experiment or misinterprets data in order to find it (see also subject-expectancy effect).
- **Context effect**
  Cognition and memory are dependent on context, in the sense that out-of-context memories are more difficult to retrieve than in-context memories.

Many large-scale accidents in industry or in transportation have been attributed to poor decisions, affected by cognitive biases such as the above. In the field of Human Factors & Ergonomics, cognitive biases in thinking and decision making are thoroughly studied to foresee such situations, and thus, design so as to prevent them.

2.3.3 Cognitive Strategies

Suppose you are working as a cashier in a retail store where lots of transactions with cash are carried out each day. In a hypothetical transaction the customer buys something that costs 8.35€ and gives you a 20€ note, so you must calculate the change that they must receive. Analytically, to calculate the change, you would follow this simple equation: "change money = paid money – item cost", which in our case is transformed into x=20-8.35⇒ x=11.65€. The next step would be to further analyze this amount into banknotes and coins aiming usually to return the least possible "items". In our example, that would be a banknote of 10€ and four coins of 1€, 0.50€, 0.10€, 0.05€ respectively, which you would hand back to the customer in order to end the transaction. Experienced cashiers, however, would follow a quite different path. They would immediately start building up the difference between the item cost (8.35€) and the given amount (20€) by depositing the actual coins and notes on the counter, starting from the smallest coin to eventually reach the target amount, i.e., |8.35| + 0.05 + 0.10 + 0.50 + 1 + 10 = 20. Why is this alternative method used by experienced professionals instead of the analytic one?

First and foremost, the analytic method requires two stages of calculations: one subtraction (20-8.35) and a series of numerical additions, while the alternative method does not require the subtraction stage. Secondly, in the alternative method, contrary to the analytic one, the progressive addition of small amounts to more round values gradually eases the cognitive effort needed for addition. As a result, the alternative method used by experienced professionals requires less cognitive effort and is less error prone.

All in all, cognitive strategies are specific methods that people use to perform cognitive tasks, including all sorts of reasoning, planning, arithmetic, etc. Typically, cognitive strategies exploit contextual or task-specific opportunities and are very often used when performing routine cognitive tasks. Such strategies help us be more effective, therefore, it is critical for a designer to recognize, understand and design likewise, so as to assist them.

Our inherent proneness to lighten our cognitive effort leads us to devise strategies to render our cognitive tasks more effective and less copious, but it may sometimes involve taking risks sacrificing thoroughness to attain faster results. As we have already seen with Hick's law, our inner urge to come to a decision fast will often supersede our need for thoroughness, even in the absence of time pressure. Such is the case with most strenuous reasoning problems that require either tedious calculations or numerous planning steps, making it possible to lead to erroneous decision making if not assisted properly.

2.4 The information processing model

The information processing model was proposed by Wickens (1992) and offers a practical way to get a grasp of the human cognitive system as a sequence of discrete functional stages. These stages are indeed identifiable, not only by experimental manipulations but also by converging evidence from brain physiology. For example, it makes sense to distinguish a perceptual stage from one involving the selection and execution of action. This is because of the morphological distinctions between the perceptual and the motor cortex.

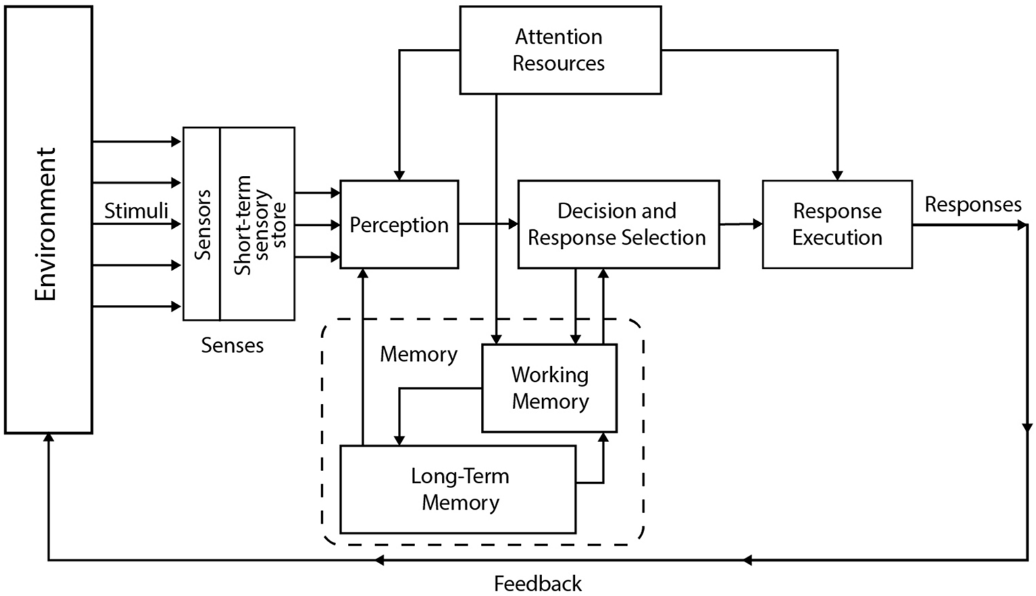


Figure 2.6. A schematic representation of the Information Processing Model.

According to this model (Figure 2.6), the basic components of the cognitive system and their corresponding functions are:

- Sensors, which collect stimuli emitted from the environment (visual, acoustic, etc.) or from the human body itself (e.g., kinesthetic).
- Short-term sensory store, in which stimuli are briefly retained in an unprocessed form before being selected and processed by perception.
- Perception, where specific stimuli are selected to be organized, identified and interpreted as information depending on their inherit characteristics (duration, intensity etc.), prior experience from LTM and the intentions and the expectations of the individual.
- Short term / working memory, in which the information is temporarily stored in an active state and manipulated (e.g., synthesized, compared, computed). This function plays a significant role in reasoning and guiding behavior and decision-making.
- Long-term memory, in which selected information perceived and/or internally processed cumulatively in the past is

organized (i), as knowledge and experience, and (ii) as long-term intentional structures of the individual.

- Attention and mobilization, which refer to psycho-mental states, with the first directing the focus of interest on certain elements of reality, and the second directing the interest in achieving certain goals.
- Decision and response selection, which refers to the executive function of choosing / deciding on ways to act.

The whole process of cognition according to the Information Processing Model can be summarized as follows: received information from the environment enters through the senses, it is then selectively perceived, subject to attentional focus and LTM, and then, it is processed in the Working Memory through cognitive mechanisms, such as comparison and, matching with existing mental structures. Through this process, an individual may make decisions and take actions to respond to external demands, or simply enrich their knowledge. In the case of active response, the senses will in turn inform the individual whether their actions have led to the desired result or, if another plan of action should be considered. This perceptual, processing and response cycle (also called action-cycle) continues until the individual considers their goal as accomplished.

An example of cognitive activity analysis based on the information processing model is the following: suppose a car driver approaching a red traffic light. Their visual system collects this stimulus in the STSS, selects it through perception and forwards it to the STM. This perceived information is signified through the knowledge of the Road Traffic Code in the LTM, and the driver decides that they should stop their car. The procedural knowledge of how the vehicle stops –again in the LTM– is activated and the muscles of the right foot are instructed to press the brake pedal. The driver's vestibular system is constantly collecting information about the deceleration of the vehicle in relation to its distance from the signal, and it processes them in the STM. Along with their experience on similar situations, the driver may need to modify the force exerted by the muscles on the foot pedal, so that the vehicle stops at the desired point, through a recurring perceptual, processing and response cycle.

2.4.1 The capacity of conscious cognitive processing

During any cognitive activity, there is a continuous flow of information between the world and the individual. A common problem in cognitive psychology has been to assess the capacity / capability of the human brain for multichannel information treatment. The question arises from the fact that humans appear to be capable of doing many simultaneous tasks in some cases but not in others. Evidently, people seem able to perform two tasks of different nature at the same time, e.g., simultaneously riding a bike on the countryside and solving equations in their mind. However, it is not clear whether the information processed in such a dual task situation can exceed a subject's single task processing capacity at its maximum sustainable rate. Two theoretical approaches have attempted to answer this question, the so-called limited capacity or "single-channel" theories (Broadbent 1958, Craik 1948, Welford, 1952) versus "multiple resource" theories (Allport 1980, Wickens 1984). The former, based on Shannon's "Mathematical Theory of Communication", maintain that two tasks cannot be performed simultaneously because they compete for a common processing resource of limited capacity somewhere in the brain, while the latter argue that parallel processing is possible in cases where two or more tasks draw on separate rather than common processing resources.

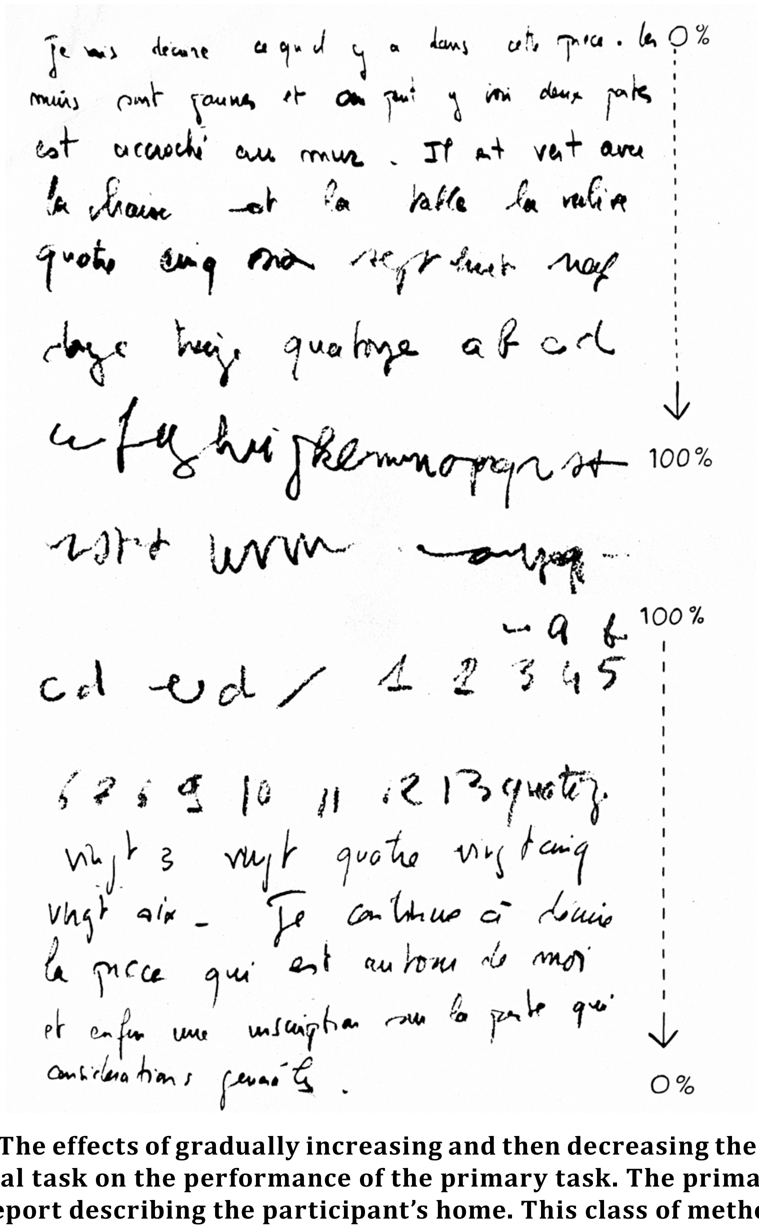


Figure 2.7. The effects of gradually increasing and then decreasing the demands of an additional task on the performance of the primary task. The primary task was writing a report describing the participant's home. This class of methods -termed "dual task"- has been widely employed in the past for assessing cognitive workload.

While numerous experiments, like the one in Figure 2.7, seemed to provide strong support for the single-channel hypothesis (Gladstones, 1989), it was unclear, until recently, where exactly the processing bottleneck occurs. More recent studies using functional magnetic resonance imaging (fMRI) revealed that such bottleneck occurs at the parietal and frontal regions of the brain, the regions where the physical manifestation of consciousness is embedded (Dehaene, 2009). Specifically, it has been shown that only one item can be present at these regions at any given time in point. However, functional neuroimaging experiments also helped reveal other relevant issues of brain functioning, such as:

1. a considerable amount of processing is possible at a subliminal (i.e., below consciousness) level, a finding partially supporting the multiple resource hypothesis,
2. attention is a prerequisite of consciousness,
3. consciousness is required for some specific cognitive tasks, including those that require durable information maintenance, novel combinations of operations, or the spontaneous generation of intentional behavior (Dehaene, 2009).

Some practical implications of these findings for the design of human-machine systems are the following:

- Where two or more genuinely independent and low redundancy tasks require continuous human attention and response, designers should not count on any capacity for parallel processing (Gladstones, 1989).
- Sharing such parallel tasks among different input and/or output modalities will not necessarily improve performance (Gladstones, 1989).
- Where the operation of high-performance systems involves the division of attention designers should use the single channel hypothesis as the first approximation (Gladstones, 1989).
- The information provided to assist decision making must be effortlessly processed in order to free cognitive resources as soon as possible, and therefore, give the opportunity to shift the attention to other important information from the environment.
- Skill development, and thus, subliminal treatment of peripheral tasks, releases attentional resources for critical tasks that unavoidably require them.

2.4.2 The limits of the information processing paradigm

The information processing paradigm and its related concepts owe much to the seminal work in cognitive psychology in the 1950s, 1960s, and 1970s, which applied the metaphor of the digital computer to human behavior. The digital computer metaphor remains until today the most pervasive way to understand human thinking even in lay terms.

The information-processing paradigm does a good service in providing structured and coherent models of human cognitive processing by identifying its various stages. For instance, different sources of workload may have different effects on the various stages. Decision-making biases can be characterized by whether they influence perception, diagnosis, and action selection or not. Besides this, the different stages may also be responsible for the commission of qualitatively different kinds of errors (e.g., errors of perception, of interpretation, of association, or errors of action selection).

The information processing approach is quite suitable for analyzing novel activities in rationally structured domains and/or "generic human subjects", where reflective reasoning dominates, and meaning/ontology generation can be sufficiently restrained by the analyst. However, its general and detached character, makes it less adequate for coping with the richness and immediacy of expert performance in real world environments. Even Herbert Simon, one of the fathers of the information processing paradigm, had noted that: "in real life, there is no well-defined, unique, and static problem but rather one which is constantly changing, whose definition is modified on the basis of information that the agents extract from their memory or they obtain through responses from their environment, to the actions that they have executed" (Simon, 1977, p.239).

In fact, the information processing paradigm has important ratifications in our conceptions of cognition, as it presupposes at least three axioms:

- The human necessitates an internal representation of external environment in the mind to be able to respond appropriately[3].
- There is a clear dichotomy between mind and body, with the mind functioning as an intelligent controller and the body as the executional automaton.
- There is a clear-cut divide between what is considered internal and what external to an organism (i.e., the environment).

[3] Newell & Simon (1972) in their seminal work on human information processing stress that intelligent behaviour presupposes an internal apparatus of representing the world. In this way, cognitive behaviour can only be explained if we assume that an agent reacts by internally representing the relevant elements of situations, in which s/he finds her/himself. Insofar, as her/his representation of the situation is correct, the agent's behaviour will be adequate.

In the remainder of this chapter, we will introduce a number of theoretical concepts that do not adhere to the above axioms, and in a way, they challenge the information processing paradigm. These concepts, stemming from diverse disciplines, such as Semiotics, Phenomenology and Neuroscience, are meant to broaden the readers' understanding of human cognition and to provide alternative models for its study, especially concerning the analysis of activity in naturalistic settings and/or of expert performance.

2.5 Semiotics and meaning-making

So far, we have used the term "information" in general terms, without referring directly to the way or the form in which it is conveyed, perceived and transformed in the mind.

But what is meant by the term "information" in human cognition? Is it the external stimulus, the filtered perception of the stimulus or the integration of perception with internal experience/expectations (i.e., its meaning)? Here is an example: if a person walks backwards, he can experience the existence of an obstacle even without seeing it, through the resistance that the muscles on his leg will encounter. In this case we have a kinesthetic[4] signal, which was perceived as resistance that has been integrated with internal experience/expectation as an obstacle. The example above illustrates the process of information transformation from the external environment to the human mind or, in other terms, the process of "meaning-making". So, although all three levels count as information, each one serves a different role.

Note that, although signals can be seen as "objective" information (i.e., only dependent on the external environment), its perception and meaning are not. Indeed, the process of meaning-making entails filtering through perception and integration with a person's experience/expectations, both of which are fundamentally subjective. This clarification is necessary, since the same meaning can be generated through many different signals, but also one signal can be the carrier of many and different meanings.

Human meaning-making is a particularly complex subject matter studied within a special branch of philosophical thought known as Semiotics. Semiotics defines the "sign" as the fundamental unit of meaning-making. In what follows we will briefly introduce some basic semiotic concepts that, although introductory and quite simplified, have been proven quite useful for design purposes.

[4] Kinesthetic are the signals transmitted to the brain by the activity of the muscles and carry information about their dynamic state.

2.5.1 Dyadic sign

According to Ferdinand de Saussure (1857–1913) the “sign” is a binary concept, consisting of a body, the “signifier”, or otherwise the “signal” and the “signified”, or to put it simply, the information that conveys its meaning (de Saussure, 1959). Thus, for example, words are “signs” where the “signifier” (signal) is their written form or their oral phoneme and “signified” the information they convey, their meaning (Figure 2.8).

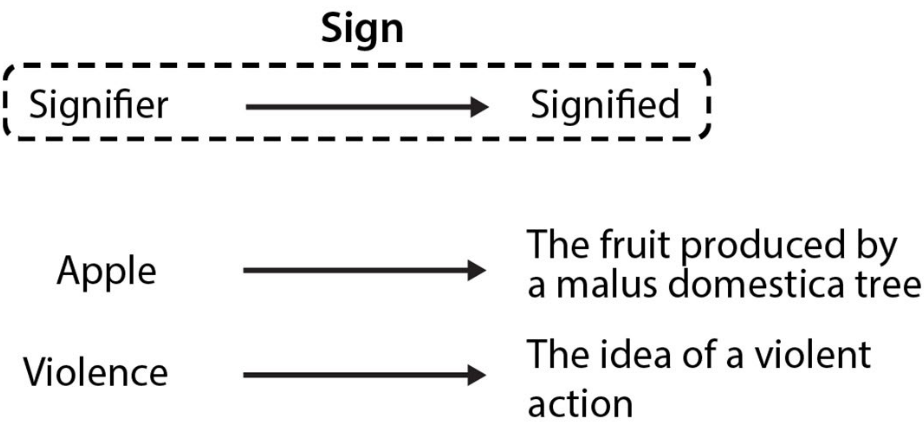


Figure 2.8 Sign, Signifier and Signified

The same idea (e.g., the fruit produced by a malus domestica tree) can be conveyed by many signifiers, whether it is the same word in different languages or an image of that fruit, etc. (Figure 2.9). A signifier may be associated with its signified in more or less arbitrary ways.

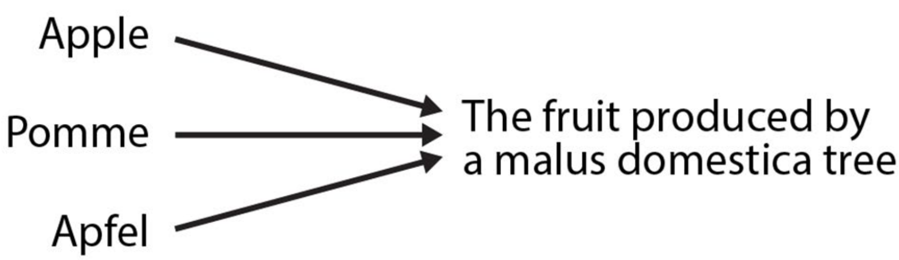


Figure 2.9. The same signified may be conveyed by various signifiers.

2.5.2 Triadic sign

In order to deal with this ambiguity, Charles Sanders Peirce’s (1839 – 1914) theory of Semiotics introduced a triadic concept which includes the “sign”, the “object” and the ““interpretant” (Deledalle, 2000). In Peirce’s Semiotics, a sign is defined as anything that is identified by (or refers to) something else (called its object) and thus, it has an effect on a person in the form of a secondary sign (called interpretant) (Figure 2.10). This effect is called interpretation. In this way, Peirce recognizes a possible objective correlation between signal and object, but according to his theory, the sign is a sign only insofar as it is, at least potentially, interpretable by a mind. Furthermore, he suggested that all thought comes as recursive sets of such triadic inferences. The result is a general theory of the production of meaning,

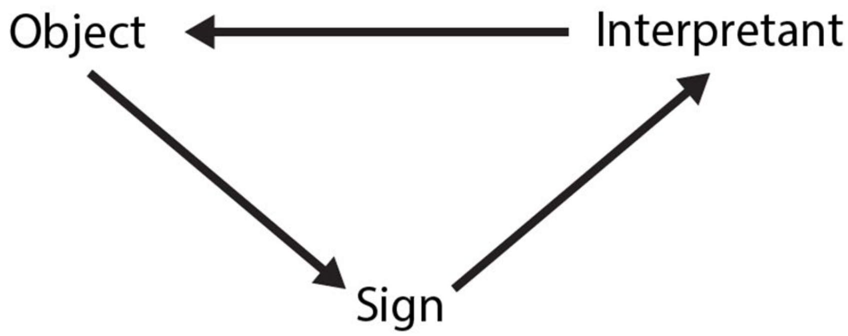


Figure 2.10. The Trinity Point of C.S. Peirce including the Object, the Sign and the Interpretant.

For example:

- A ”Left Turn” traffic plate refers to a natural turn of the road and is interpreted by someone as a warning, if he knows the specific graphic design protocol.
- The “No Parking” traffic sign refers to an administrative regulation and is interpreted by someone as such, as long as he knows the specific graphic design convention and the administrative regulation behind it.
- The sound of a school bell refers to the entrance or exit of the classroom and is interpreted as such by the students and teachers at that school. Respectively, the specific sound of a doorbell in a house refers to a request for a person to enter the house and is interpreted as such by someone who knows the specific sound.
- A specific cloud formation happens due to a disturbance of atmospheric pressure and is interpreted as an impending storm by a meteorologist or a farmer.
- A specific alkaline odor in an industrial refrigerator refers to the presence of ammonia vapors and is interpreted as a leak from the workers in the area.

The above examples are used to show that:

- signs may or may not be designed for a specific purpose,
  the objects to which they refer can be natural phenomena, social conventions, or other signs,
- interpretations do not necessarily coincide with objects but are shaped by the sign and the particular characteristics of the human receiver at any given time.

2.5.3 Typology of signs

In contrast to Saussure, who approached semiotics from a study of linguistics, Peirce, considered "words" to be just one particular type of sign and extended the concept to encompass any mediational means to understanding. Besides words and symbols, he considered as signs any semblance or any type of indicator. Peirce distinguished three basic types of association (Figure 2.11):

- **Symbolic**: the signifier is purely conventional and does not share any property with the signified; it is arbitrary. The symbolic connection must always be taught (e.g., the word "stop", the red light, number 7). On the other hand, precisely because of its arbitrary nature, a symbolic connection may convey a multitude of meanings, as is often the case with the words of a language.
- **Iconic**: the signifier imitates the signified resembling it in some way, such as a portrait, a blueprint, a model or a mimetic gesture. Depending on the fidelity of the imitation, the correlation between signifier and signified may be direct or require some familiarity.
- **Indicative**: the signifier is connected in some concrete way (natural or teleological) with the signified. This connection can be observed or inferred (e.g., a footprint in the snow with the passing of someone; the smoke with the fire; the mercury level of a thermometer with the temperature; the spirit level's bubble with the slope of a surface; a knocking on the door with someone behind it). The indicative signifiers are either immediately perceived or constructed through personal experience.

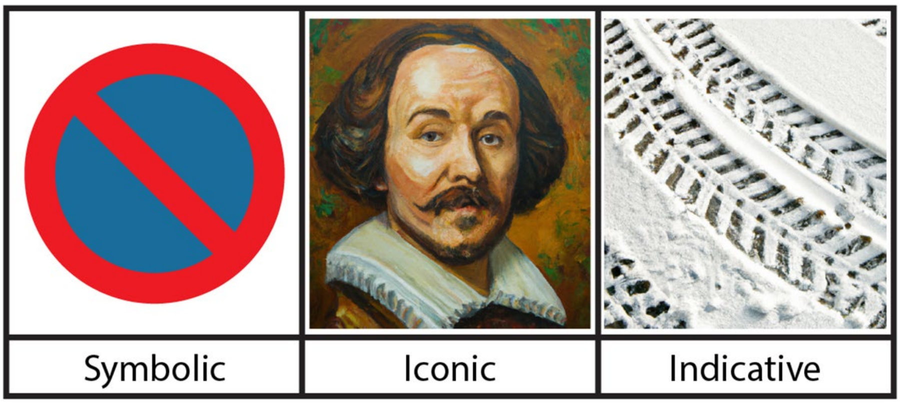


Figure 2.11. Tree basic types of signifiers.

Following the above, the signifier of a gesture is the specific movement and shaping of the hand and the signified a culturally developed statement.

Thus, when we form the V-shape with the index and the middle finger of our hand (Figure 2.12), it can either mean the quantity "two" (iconic connection), or the meaning of "victory" (symbolic connection), or again a statement of insult or discredit for an Anglo-Saxon (symbolic connection). It becomes clear that, in practice, the signifier-signified relationship is drastically influenced by the knowledge and previous experiences of the receiver, as well as the cultural context in which one belongs and, finally, by the situational context in which a person is at that specific moment.

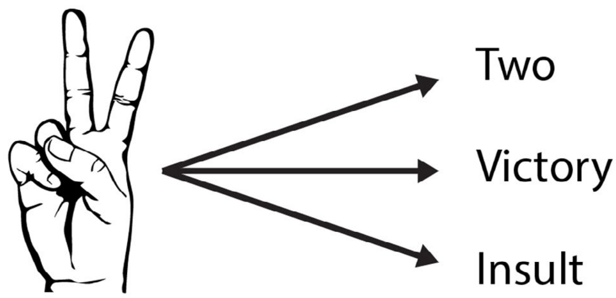


Figure 2.12. Different meanings of the V-shape gesture.

2.5.4 Practical implications for design

For design purposes it is important to note that signal/signifier characteristics should be decided based on representation heuristics and ease of interpretation by the intended group of users. Therefore:

- When an object to be conveyed can easily be represented visually, it is advised to use iconic signifiers (e.g., garbage bin, mailbox, fuel level).
- When a sign system contains many distinct objects and/or abstract objects, with no straightforward iconic representation (e.g., pressure level, danger, cease of operation-turn off – shut down), a designer should consider symbolic signifiers.
- When a sign can easily be implemented based on the natural phenomenon it represents, it is advised to use indicative signifiers (e.g., a windsock for wind direction and speed).
- When accuracy is needed as in objects with parametric values (e.g., speed, pressure), a symbolic signifier should be considered.
- When directedness of interpretation is more vital than accuracy, then an indicative or iconic signifier should be preferred.
- When standard conventions (either iconic or symbolic) already exist, these should be followed as much as possible (e.g., shopping cart icon in e-commerce platforms).

In any case, a designed sign may deliberatively embody more than one modality. For example, a windsock may contain indexical and symbolic elements (Figure 2.13), a public toilet sign may contain a symbolic element (i.e., the initials "WC"), an iconic one (i.e., man / woman icon) and an indexical one (i.e., two adjacent doors); a fuel level sign may contain both a graphical depiction of the fuel tank with iconic fuel level and a numerical value for the percentage of fuel remaining. Such complementarity and redundancy in signifier elements may prove advantageous if judiciously designed and not overloaded. Typical examples of complementary signifiers are traffic signs where the color scheme and shape or the sign contain a symbolic meaning (e.g., yellow triangle meaning danger), and an animal pictogram (e.g., a deer) specifying the particular danger being a wild animal crossing the road.

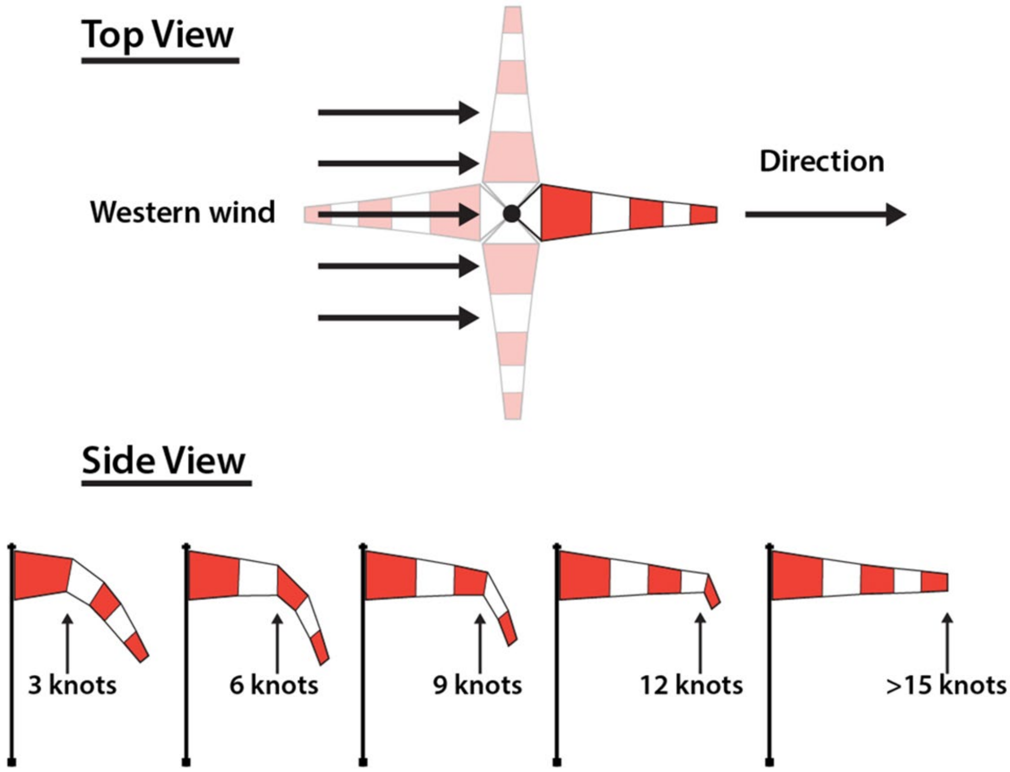


Figure 2.13. A windsock is an indicator of wind direction and speed. When pointing due East, it indicates a western wind. Wind strength is indicated by the length of the windsock which remains horizontal. These qualities can be read by all observers as they are indicative signs. Each alternating orange or white stripe adds 3 knots to the wind speed, a symbolic notion that gives more reading accuracy to the observers who know this convention.

Semiotic concepts are not only relevant for design purposes but also for observational ergonomics studies. As already mentioned, signs are not always deliberately designed for a designated purpose. They might, also, come naturally as indexes or be discovered to function as such through experience. In fact, elaborate use of “hidden” or non-obvious signs is what distinguishes the expert from a novice in many professions.

For this purpose, Montmollin (1974) proposed some practical sign distinctions that help an analyst in revealing the mental activities and competences of experienced workers. Two of these distinctions are presented below:

Formal vs. informal signs. Formal signs are those designed by humans specifically to convey information. Example: a thermometer emits a standard sign that informs us about the temperature of a solid, liquid or gas. The information transmitted by a formal sign is usually coded and quantified. Informal signs, on the contrary, are not specifically designed to transmit information but since they do so anyway, this information can be used by someone when performing a task. Example: the red color of molten steel in its smelting furnace is for an experienced worker an informal sign for the temperature of the metal. Also, the characteristic noise emitted by a machine can be an informal sign informing a worker about the need to lubricate the machine. The use of informal signs is an integral part of the mental skills of experienced workers.

Some of the most subtle informal signs are the kinesthetic ones. Example: A metalworking file operator adjusts his movements to a large extent based on the information provided by the muscles of his hands about the resistance of the workpiece. At the same time, the same operator regulates his activities based on the visual signals he collects from the piece which inform him about the course of the process. When designing technological devices, kinesthetic signs are often more efficient when used properly, than signs transmitted to our other senses, as they are less likely to be omitted or misinterpreted.

Explicit vs. implicit signs. The explicit signs are the ones that people use consciously and which they can report spontaneously, once asked. Implicit signs on the other hand are those used unconsciously. Implicit signs are almost always informal and require specialized techniques to be discovered.

2.6 Ecological and embodied approaches to cognition

In contrast to the information-processing paradigm, the ecological paradigm to human cognition suggests that, in functional terms, mental processes often extend beyond the brain, to encompass the body and even the external environment. As a result, in order to understand intelligent behavior, one needs to consider physical action and proximal environment as integral parts of the cognitive system.

> *"First, rather than analyzing the distinct stage sequence of information in the human mind, the ecological approach puts emphasis on the integrated flow of information through the human as a whole. Second, it emphasizes the inexorable embeddedness of humans in the physical and cultural environment. Accordingly, it focuses heavily on modeling the perceptual characteristics of the environment to which the human is "tuned" and responds in order to meet the goals of a particular task. Third, in the ecological approach, bodily action and perception are closely linked, since to act is to change what is perceived, and to perceive*

is to change the basis of action in a manner consistent with the closed-loop representation" (Wickens & Carswell 2021).

As a case in point regarding the division among mind, body and environment, read the following famous extract from Gregory Bateson's book "Steps to an Ecology of Mind".

Figure 2.14. The stick of the blind man as an extension of his body.

"Suppose I am a blind man, and I use a stick (Figure 2.14). I go tap, tap, tap. Where do I start? Is my mental system bounded at the handle of the stick? Is it bounded by my skin? Does it start halfway up the stick? Does it start at the tip of the stick? But these are nonsense questions. The stick is a pathway along which transforms of difference are being transmitted. The way to delineate the system is to draw the limiting line in such a way that you do not cut any of these pathways in ways which leave things inexplicable. If what you are trying to explain is a given piece of behavior, such as the locomotion of the blind man, then, for this purpose, you will need the street, the stick, the man; the street, the stick, and so on, round and round. But when the blind man sits down to eat his lunch, his stick and its messages will no longer be relevant—if it is his eating that you want to understand." (Bateson 1972, p.459)

Indeed, in situations of effortless coping, the coupling of the mind with the body and the environment is so strong that an analysis in terms of mental stages and input – output becomes unproductive. The environment itself becomes its best representation, and physical action is directed as much from conscious thought processes as it is from directly perceived environmental opportunities.

As a consequence of these properties, the ecological approach is essential to describing human behavior in interaction with the natural environment (e.g., driving or manipulating objects directly) or in effortless interaction in a familiar setting (e.g., people working in domains and systems about which they are experts).

An example in point is table tennis expert play. Table tennis athletes do not consciously measure the speed and trajectory of the incoming ball, nor do they consciously manipulate their hand and racket to respond. In a way, it is as if the ball trajectory, movement of eyes, and hand-racket form a transient closed loop that functions independently from conscious mental processing. In fact, expert players may focus their conscious thoughts on tactical or strategic decisions and "let their body play".

At a theoretical level, the Ecological / Embodied approach to cognition goes well beyond the immediate material environment and the physical body. Jacques Theureau (2002) a prominent cognitive ergonomist and proponent of the ecological / embodied approach proposes to adhere to the following theoretical hypotheses when studying human activity in context. According to Theureau (ibid) human activity should be considered as:

autonomous, i.e., consists of asymmetrical interactions between the acting human and his/her environment, in the sense that his/her interactions concern not the environment as an observer from the outside could apprehend it, but his/her "proper ecology", i.e., what, in this environment, is relevant for the internal structure of the human at every instant,

cognitive, i.e., manifests and continually develops knowledge (i.e., manipulation of symbols and development of mental representations),

embodied, i.e., consists of a continuum between cognition, physical action, communication, and emotion,

dynamically situated, i.e., always appeals to current resources, which stem from constantly changing material, social, and cultural circumstances,

indissolubly individual and collective, i.e., even individual events are interwoven with collective events,

cultured, i.e., inseparable from a cultural situation that is either collectively shared or individually to various degrees, and finally,

experienced, more precisely causing experience (i.e., awareness / meaning / feelings) for the acting human at every instant, however partial and fleeting this might be.

2.6.1 The notion of embodiment

Perceiving – acting

The habituated scanning of my eyes and the movement of my hand when trying to accomplish an action in a specific computer application are important. Peripheral sight and hand trajectory over the mou se in order to print a file are important. It forms an integral part of knowing how to print a file. Hearing of the printer noise is also important as a subtle verification that all goes well. Response lags on a specific PC or while at a specific Internet site are important. I usually neither think about them nor measure them, I just coordinate my hand and eyes movement with these lags to perform my actions with the least tension. All the above bodily patterns are an integral part of coping with the world that the body "learns".

*A while ago, a friend asked me to go to her house and feed her dogs while she was away. As the house had an alarm, she had sent me a message with the alarm code (#1735). Getting there it turned out that the code didn't work, so I called her up. She insisted that the code was correct, and she had been using it every day for the past 10 years. Failing to deactivate the alarm with a second try I left without being able to accomplish her request. It was not until she returned home the next day and stood in front of the keypad that she realized what the problem was. Her hand went straight to the left bottom corner dialling "*1735" as the correct code had in fact an asterisk (*) instead of a number sign (#). Her realization came only while acting on the keypad and not while reflecting on the symbolic notion of the code (Figure 2.15).*

Figure 2.15. Remembering through action; a security code is embedded as much in the mind as in the hands of its holder.

So where does the knowledge of the alarm code lie in the above example? Is it purely a mental symbolic representation embedded somewhere in the brain? According to the ecological paradigm, the existence of a real keypad is an integral part of the knowledge or retrieval process. Not only its visual image, but its material presence and the physical act of dialing are integral to remembering the code. Why? Because this is how she learned it, not by rehearsing it mentally, but by physically acting on the keypad time and time again. Such knowledge is in one word "embodied".

In theoretical terms, embodiment refers to the process of 'owning' or 'internalizing' our actual experience of things. It can be defined as the mind-body skill of situational discrimination and seamless immediate action. It draws from the phenomenological tradition of European philosophy and particularly from the work of M. Heidegger (Dreyfus 1991), and in a more radical way from the work of Merleau-Ponty. According to the latter, in everyday, absorbed, skillful coping, acting is experienced as a steady flow of activity in response to one's sense of the situation. Accordingly, human behavior can neither be explained in a behaviorist way in terms of external causes, nor internally in terms of conscious intentionality. Rather, it had better be explained structurally, in terms of the physical structures of the body and nervous system as they develop in a circular interplay within the world. The world does not determine our perception, nor does our perception constitute the world. As Merleau-Ponty, cited by Dreyfus, (1996) puts it: “The relations between the organism and his milieu are not relations of linear causality but of circular causality”.

Proponents of the embodied perspective reject the idea of cognition as solely processing of abstract symbolic representations in the brain. They emphasize on the ways skilled workers offload parts of the representational and processing burden of cognition in the external world and in the motor and perceptual subsystems that interact with it. Skill development in the above sense is precisely the progressive diffusion of cognitive processing to the environment and to the body. This diffusion results in fluid action referred in the literature under various names such as “skill-based behavior” (Rasmussen et al., 1994), “readiness-to-hand” (Winograd & Flores, 1987) or “operationalization” (Nardi, 1996).

Following the above, an expert Computer Aided Design (CAD) draftsman, when immersed in his multi-layered CAD world, he does not “reason” in terms of layers, but “acts upon” layers with his eyes and hands (Goel 1995). Much alike, stock market traders, before reflecting, perceive financial opportunities through the configuration of tables or superposition of graphs. (Nesbitt, 2001).

Such pre-reflective flawless action extends well beyond the physically observable to encompass cultural and linguistic interactions (Mingers, 2001). This can be observed with expert developers working on command line interfaces. From a phenomenological perspective, such individuals neither type commands nor read outputs from the computer screen. The object of their concern lays elsewhere; it is mirrored in spontaneous verbalizations such as “I am trying to locate this bug” or “there is some .dll conflict”. The speed and seamlessness of interaction with the operating system, coupled by simultaneous statements such as the above, indicate that the specific interface has been embodied to a large extent.

Imagine the enormous difficulty many of us would have if asked today to write a report in a typewriter. While working through a word processor, we are seldom aware that we exploit features like cut – copy – paste, find-replace and the like, not only as facilitators for speed or text formatting but also as an integral part of how we do the job. We do not merely put our internal thoughts into words, we construct our thoughts visually and reflectively inside the word processing environment.

A tale on oyster foraging, or how phenomenal ontology is generated by directly coupling environment perception and action.

Figure 2.16. A seabed landscape is just a colorful and complex landscape for the untrained eye.

A seabed landscape is just a seabed landscape unless you are an oyster catcher. Then, it becomes the seabed landscape of the oyster catcher (Figure 2.16).

One might discover how to spot oysters by himself, or he might be instructed how. In any case, it is done by observing one in its natural environment. Descriptive ways of instruction are usually hopeless (everybody knows how an octopus looks like, but only octopus catchers can spot octopuses on the seabed). The oyster might have a more or less "certain color", a more or less "certain size", a more or less "certain shape", but hours of descriptive instruction are less helpful compared to pointing on a single one on the seabed.

When I am down there, in search for them, at first, I see a lot more or a lot less that there actually are (a lot of false alarms). Progressively, I learn to better filter the perceptual field up to the point of actually faultlessly spotting at first glance a "good sized oyster", from a distance of 5 to 10 meters. As my immersion into the oyster environment progresses, I start to orient myself to the right parts of the seabed. I look down at the seabed and think "this looks like a good rock for oysters". I dive down and usually I find some. To the untrained observer this is just a seabed, to the oyster catcher it has become an oyster breeding environment. In contemplating, later on, this particular "acquired skill" I may state that:

- *I scan the seabed searching for a particular color, when I spot it,*
- *I check the size and shape of the thus colored spot,*
- *I check the surroundings ...usually oysters grow on southerly slopes (I think they need the sun) ...usually oysters grow along with this or that type of flora.*
- *The sea current seems important to them etc.*

If I contemplate or if asked by somebody, I will probably build a micro theory such as above. But really when I am down there in search for *them ...I just orient myself in this environment and see (recognize) them. I am reflecting about none of the above conditions. In fact, if and when I try to reflect on such rules, while still an apprentice, it fast becomes a mess... overwhelmingly complex and unproductive.*

The above examples exemplify the need to acknowledge that people progressively establish authentic ecologies, through engaged action in the world. They should not be viewed as mere humans but as encultured individuals with an embodied understanding of their activities.

As stated by Winograd and Flores (1987), objects do not exist independently of what we do. It is what we do that constructs the objects of our concern, and this can happen only through involved embodied action. Activity is constitutive of ontology, not independent of it.

Interactive systems design is, therefore, first and foremost ontological; it starts by defining codes, categories, meaning structures, properties, relations, states of affairs, events and the like. These are the core material for conceiving any interactive system, before tackling surface features such as visual interfaces or input means. In this sense, the ecological and embodied approach to cognition prepares the designer to look through the eyes of experienced individuals so as to elicit their embodied understanding and the objects of their concern.

2.6.2 Tools as integral components of cognition

Tools have a profound effect on cognition. They do not just make our tasks easier, but more importantly, they may alter the nature of the task itself. As the use of a shovel alters not only performance but the movements of the human body, in the same way an abacus or a digital calculator alters the nature of the cognitive processes being employed by the person performing the calculation task. When studying cognitive activity in naturalistic settings, one should necessarily include to the cognitive system the tools being used, as these fundamentally alter the cognitive processes needed for task accomplishment.

Consider someone assigned to count the number of bags passing in front of him. If the pace of passing bags is rather slow, the obvious thing to do for most humans, apart from loud or silent oral uttering of consecutive number names, would be to use their fingers as a physical memory aid. If the number of bags increase beyond a certain number, the “fingers” trick would not be enough. Then the person would probably try to offload their memory by inscribing it on the environment (Figure 2.17).

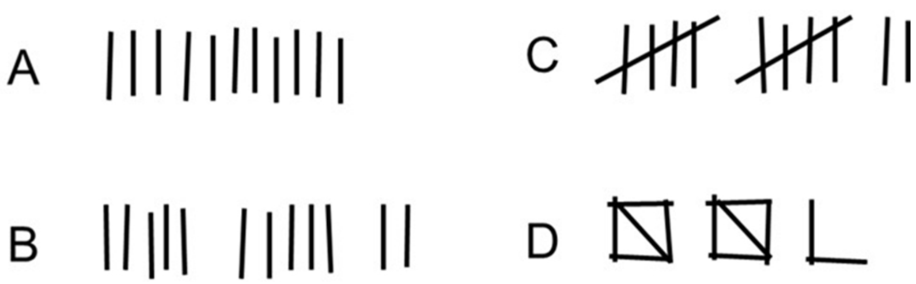


Figure 2.17. Alternative ways to inscribe counted items on the environment.

This well-known trick (i.e., the tally mark) is, in fact, a primitive cognitive tool. In version A, it offloads the user from the tedious and error-prone task of continuously keeping in memory the ongoing count, while in versions B, C and D, it facilitates intermediate and final valuation by greatly reducing the required mental effort and possible final counting errors.

Such offloading of cognitive work, occurring due to the invention of artifacts, is pervasive throughout the history of human civilization. Up to the Middle Ages, mariners used to determine how fast their ship was moving, by throwing a floating object over the side from the ship's bow and then counting the amount of time elapsed before the object passed the stern. Depending on the ship's length and the pace of oral counting they could, thus, have a rough indication of the ship's speed, and thus the estimated time of arrival at a certain port. If, for instance, the floating object passed through the ship's stern at count number 5, it meant that the speed was double than when the floating object passed after counting up to 10. This measurement was obviously ship-length specific, but depending on experience, it allowed fair predictions of trip duration.[5]

From the 15th century on, mariners started using a more accurate method. A triangular piece of wood (called "chip") was attached to a rope and was thrown behind the ship as a drogue. As the vessel moved forward, the line of rope was allowed to roll out freely for a specific time interval with the aid of an hourglass. The sailors would then retrieve the line and measure the distance travelled in "fathoms" (an anthropocentric measurement unit originating from the span of a man's outstretched arms, approximatively 1.85m). Through appropriate fathom/timetables they would then derive the vessels speed in Fathoms per hourglass (time) units.

[5] Note that the estimation result in elapsed time, although rough, was directly proportional to the main question it was used for i.e., how long will it take to get somewhere (e.g., to sail from Malta to Crete with a measured "6" elapsed time it takes around three days, while with a measured "9" elapsed time it takes around four and a half days). This is in contrast to speed measured in distance / time units which, to answer the same question, even though more accurate and universal, needs a division (i.e., Distance / Speed).

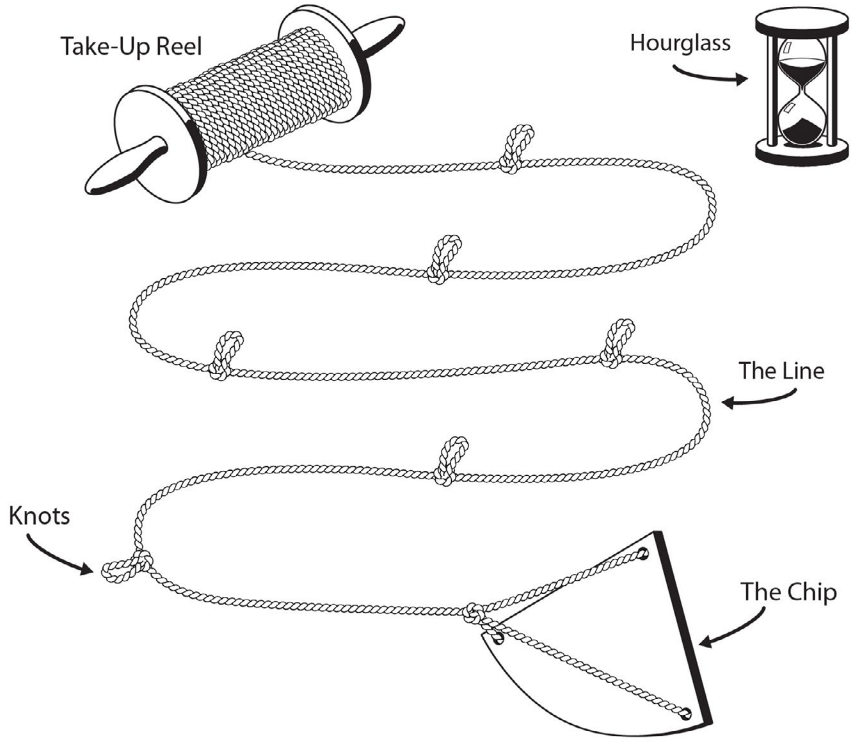


Figure 2.18. A chip thrown behind the ship as a drogue to measure speed. Physical knots at appropriate distances on the drogue line signify speed in knot units (nautical miles/ hour).

Later, sailors began to mark the rope line by tying knots at fixed equal intervals (Figure 2.18). Afterward, the number of knots that had passed over the ship's stern was counted and used in calculating the vessel's speed, doing away with the physical task of measurement and its inherent variability. With the spread of the nautical mile (i.e., 1852m) as a standard unit of distance measure at sea in the 16th century, by proper adjustment, one knot was calibrated to signify one nautical mile per hour (most usually 8 fathoms between consecutive knots and roughly 30 seconds of time interval). Therefore, 6 knots passing over the stern in 30 seconds meant that the ship was travelling at a speed of 6 nautical miles per hour. Eventually, the physical knot on a line of rope lent its name to the term "Knot" used in contemporary navigation to designate Nautical Miles / Hour.

The story above describes in simplified terms the historical development of a tool or otherwise, a cognitive aid. By progressively offloading physical and cognitive subtasks of the human sailor to a material artifact (in this case the calibrated knotted line and hourglass), estimation of a ship's speed became less error prone, more accurate, communicable and effortless. Note, also, that the knowledge and calculations needed to perform the speed estimation task were progressively embedded in the tool, thus rendering the task straightforward even for a novice user.

2.6.3 Cognitive aids: a double-edge sword

In 2007 and 2012, two major accidents happened at sea, involving two modern cruise–ships - MS Sea Diamond and Costa Concordia. Both ships sank after running aground, the first near the coastline of the island of Santorini, Greece and the second in Giglio Island, Italy. Both had come too close to charted reefs, exceeding the safety depth limits of their draft. Such an oversight from their experienced captains seems at least strange, since these ships were equipped with advanced Electronic Chart Display and Information Systems (ECDIS) that, in theory, should prevent such accidents. Paradoxically, though, ECDIS might be part of the reason for their occurrence.

Prior to the implementation of Global Positioning Systems (GPS), traditional sea charts and portolans were plotted using various optical instrument measurements of shore features, or celestial bodies. Their accuracy was, thus, constrained by the methods of data collection and also, by the lithographic processes and plotting techniques used, as well as the symbolization of features (e.g., line widths).

Used in navigation for hundreds of years, these paper charts were still limited by their medium to an accuracy that did not exceed the accuracy of cartographic processes used by the chart maker (scales between 1:5000 and 1:100000). That way, the inherent uncertainty induced by the size and resolution of the paper chart (Figure 2.19 left), as well as by the less accurate means of navigation available on board, would force the mariner to keep a safe distance from navigational hazards. The situation is now reversed; with the integration of GPS signals, mariners, nowadays, can obtain a more accurate position than the one used to compile the charts they use (Figure 2.19, right).

This may lead to maneuvering overconfidence; for example, bridge officers, to save sailing time, relying on their GPS may pass closer to hazards depicted on charts than it is prudent. Moreover, chart digitization through vectoring provides the ability to zoom-in at any given spot of the map. However, after transforming original data into mathematical curves, the resolution may falsely seem infinite. Zooming may, thus, lead mariners into taking risks without even knowing it, by exceedingly trusting their monitoring displays. As a consequence, there is an increase in incidents of navigating at the limits of a ships' safety envelope.

In reality, the digitization process of the maps was done mostly by the hand of a land surveyor/cartographer, on the basis of geospatial information collected using different techniques at different times (e.g., scanned printed maps or satellite images), the detail of which was also constrained by the accuracy and resolution of the original source.

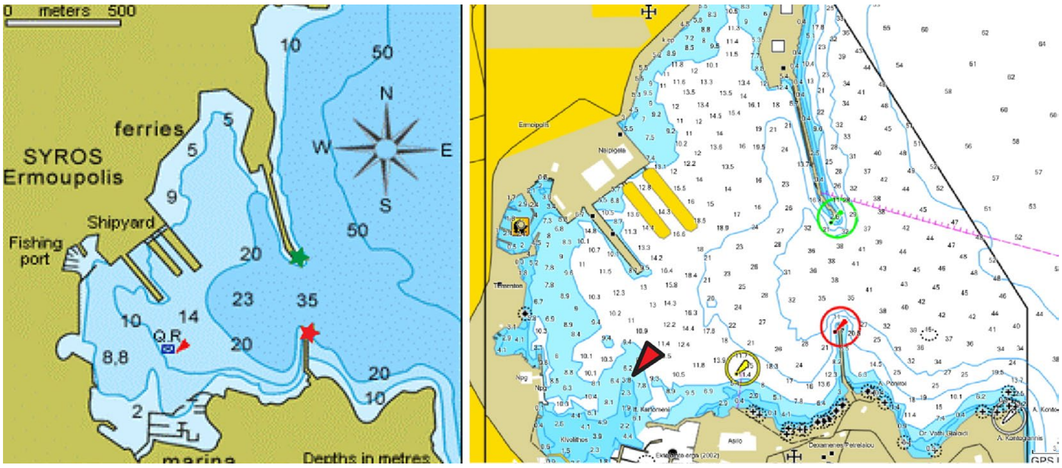


Figure 2.19. A typical paper-based navigation chart (left), the same chart area in digital form conveying detailed information on environment morphology and integrated vessel position (right).

Therefore, while modern equipment is much more accurate and precise, offering high resolution monitors, there may be a mismatch between their resolution capabilities and the underlying accuracy of the digitized maps. This leads to false overconfidence of systems' accuracy, with actual error margins being much higher than communicated to the end user. Lesson to be learned: mediated direct perception is tricky because it conveys a feeling of certainty (of actually being there) that may not be warranted by the underlying technology. This example leads to a more general caution concerning iconicity with direct relevance to modern iconic technologies (e.g., Virtual Reality): iconic representations, tend to subconsciously make us to believe that there are no gaps between reality and its representation. Hence the paradox of high-fidelity iconic representation: it may deceive us most when we think it works best.

References

Baddeley, A., and Hitch, G. (1974). *"Working memory," in The Psychology of Learning and Motivation,* ed. G. H. Bower (New York: Academic Press), 47–89.

Bateson, G. (1972). *Steps To An Ecology Of The Mind*. New York: Ballantine.

Blakemore, C. (1988). The Mind Machine. (BBC Classics). BBC Books.

Deledalle, G. (2000). Charles S. Peirce's *Philosophy of Signs: Essays in Comparative Semiotics.* Indiana University Press.

Dharani, K. (2015). The Biology of Thought, Chapter 3 - Memory, Pages 53-74, Academic Press, ISBN 9780128009000

Höfle, M., Hauck, M., Engel, A. & Senkowski, D. (2010). Pain processing in multisensory environments. *e-Neuroforum*, 16(2), 23-28. https://doi.org/10.1007/s13295-010-0004-z

Loftus, E. F. (1996). *Eyewitness testimony.* Harvard University Press.

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General,* 118(1), 100–104. https://doi.org/10.1037/0096-3445.118.1.100

Loftus, E. F. (2013) *The Myth of Repressed Memory: False Memories and Allegations of Sexual Abuse.* St.'Martin's Press.

Loftus, G., Dark, V., & Williams, D. (1979). Short-term memory factors in ground controller/pilot communication. *Human Factors*, 21, 169–181.

Gladstones, W. H., Regan, M. A., & Lee, R. B. (1989). Division of attention: The single-channel hypothesis revisited. The Quarterly Journal of Experimental Psychology, 41(1), 1-17.

Rasmussen, J., Pejtersen, A.M. & Goodstein, L. (1994). *Cognitive Systems Engineering*. John Wiley & Sons, New York.

de Saussure, F. (1959) [First published 1916]. *Course in general linguistics.* New York: Philosophy Library. ISBN 9780231157278

Simon, H. (1977) *Models of Discovery: and other topics in the methods of science*. Dordrecht, Holland: Reidel.

Stillings, N., Weisler, S, Chase, C., Feinstein, M., Garfield, J. & Rissland, E. (1995). *Cognitive Science: An Introduction, 2nd edition*, Cambridge, MA: MIT Press.

Ten Berge, T., & Van Hezewijk, R. (1999). Procedural and declarative knowledge: An evolutionary perspective. *Theory & Psychology*, 9(5), 605-624.

Theureau, J. (2002). Dynamic, living, social and cultural complex systems: principles of design-oriented analysis. *Rev. d'Intelligence Artif.*, 16, 485-516.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–101) New York: Academic Press.

Wickens C. D. (1992). Engineering psychology and human performance (2nd ed.). HarperCollins.

Wickens, C. D., & Carswell, C. M. (2021). Information processing. *Handbook of human factors and ergonomics,* 114-158.

Wisner, A. (1982). Physiologie du travail et ergonomie. Paris : Conservatoire National des Arts et Métiers. References

Chapter 03:
Models of Human Activity

Chapter 3: Models of Human Activity

Chapter Summary

Following from the basics of human cognitive functioning, in this chapter, a selection of applied models and relevant concepts are introduced that are widely used in Human Factors / Ergonomics for design and intervention purposes. The chapter begins by introducing mental models i.e., how people functionally organize knowledge for a particular domain in their mind. Subsequently, two applied models of cognitive activity when interacting with artifacts or systems are introduced, namely Donald Norman's human action cycle and Jen Rasmussen's Skills Rules Knowledge taxonomy of human activity levels. Typical human abilities and deficiencies are discussed in this context along with design principles meant to enhance seamless human – system interaction. The chapter ends with an introduction on the analysis of complex cognitive tasks typically encountered in professional domains such as diagnosis and troubleshooting and on the particular design challenges that result from technological and organizational complexity.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of previous chapters.

3.1 Mental models

At the Fiumicino Airport in Rome, a passenger, while heading to the departure Gate nº 8, comes across the signage shown in Figure 3.1.

Figure 3.1. Signage at the Fiumicino Airport in Rome when a passenger is heading to the departure Gate nº 8.

Although this specific gate number is absent from the signage, s/he will probably choose the staircase on the right assuming from the given information that the left side hosts all the odd numbered gates, while the right side hosts the even ones. This assumption manifests a tentative prediction of gate topology based on basic mathematical categorization (i.e., odd numbers to the left, even numbers to the right). After ascending the stairs, s/he will come across the signage shown in Figure 3.2.

Figure 3.2. Signage at the second floor of Fiumicino Airport in Rome.

Gate nº8 is again missing from the signage but now the numbers mentioned (7-10) do not correspond to the previous assumption, so the passenger is probably confused trying to decide if s/he is on the right pathway. The lack of grouping consistency among the different signs has caused her/his tentative assumption to collapse and, therefore, triggered the need to form a new one so as to come to a decision.

Making such assumptions is something quite common in our everyday life. The information we get from our surroundings is never complete or clear-cut, so we constantly fill-in gaps through existing knowledge to make decisions or predictions. In other words, we are constantly constructing "mental models" of the things we observe, either consciously or not.

3.1.1 A pragmatic definition of mental models

Mental models are, therefore, internal knowledge constructs offering a functional representation of the individual's understanding of its surroundings. These 'small-scale models' of external reality are progressively formed through experience and are projected onto subsequent experiences, often unconsciously, serving as a guide to interpret novel situations. It is worth noting that, whereas mental models form the basis of how we understand and interact with the world, they are by no means complete, or accurate representations of reality. They are pragmatic solutions for dealing with complexity. More often than not, they help us make successful predictions, but at times may also betray us.

A useful analogy to help illustrate the concept of mental model is a schematic diagram (Figure 3.3). Schematic diagrams are purposefully oversimplified, often distorted representations of real systems, designed to convey some specific system information, omitting all details that are not relevant so as to make this essential meaning easier to grasp. In the same manner, a mental model is a concise, often distorted, representation of reality. Adding detailed information to a mental model does not necessarily make the model more useful. Indeed, the conciseness of mental models may be the key to their utility, enabling faster intuitive decisions.

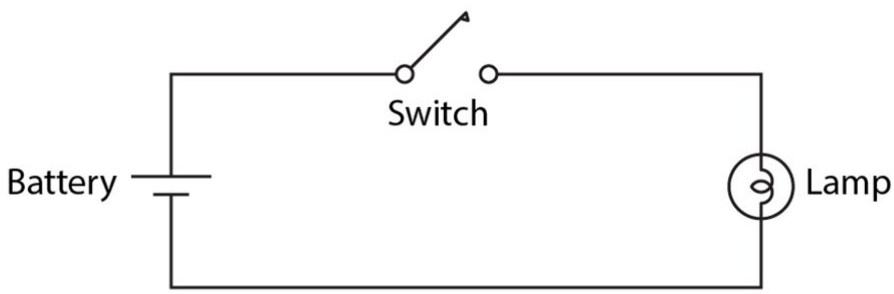


Figure 3.3. Schematic diagram of an electrical circuit; note that the diagram is intentionally simplified and distorted to convey only the basic functioning of the circuit.

Mental models are an integral part of our long-term memory, gradually created by filtering incoming information subject to our engagement/ interests and prior knowledge. As we have discussed earlier in the "long-term memory" section through the example of the bicycle, the filtering of incoming information depends on the following factors:

- its functional usefulness,
- the values and goals of the receiver,
- his/her cognitive ability,
- his/her psychological condition,
- possible established prejudices,
- the context of use at any given time.

Indeed, eye tracking experiments show that our attention is drawn differently at a scene, depending on the factors above, creating very different eye-fixation maps among different people. So, the mental models we create are not only based on already selective information but are also subject to constant enrichment or reconsideration depending on their present use. The main reason of these modifications is, on one hand, the need for cognitive economy that leads to selective focus and processing of information, and on the other hand, the inherent urge for semantic integration to produce meaning when needed. These models are therefore dynamic incorporating, as the case may be, not only the "image" of an object of concern but also its operating principles, its potential uses or even our skills with that.

A brief definition of mental models (or mental representations, cognitive representations, mental images) is that they are a hypothetical way of synthetic recording and representation in the mind of one person's knowledge about the objects around him/her, the events s/he experienced, but also his/her past actions (Denis, 1979). Mental models are gradually synthesized and can change with the enrichment of the individual's experiences.

3.1.2 How mental models assist cognition

Mental models are considered to assist cognition in several ways, the main of which are:

- They help our perception. If a person faces an X2 condition that has common features with a previously experienced X1 condition, the new X2 condition will be perceived in a similar way to X1. For example, a person familiar with driving a car with a dashboard such as the one depicted in the left-hand picture of Figure 3.4, will readily feel confident gathering information from a dashboard such as the one in the right-hand picture of the same Figure – even though they have never seen it before. Moreover, our mental models not only help but also guide our perception. As they form the basis of our expectations, they filter out environmental stimuli, perceived elements and their relationships, ignoring non expected ones. Thus, even the gaze path and the fixation points where the eye stands to gather information, when observing a scene or an object, can be explained through our mental models.

- They guide action, depicting the empirical and any theoretical knowledge that a worker has about his/her work system. In other words, according to his/her mental models, s/he determines the need for action and also, the type of actions that s/he should perform each time.
- Based on mental models, a person can predict situations, the course of events or the results of his/her actions. Example: The driver of a car can predict that turning the steering wheel to the left will turn his car to the left, even if he ignores how his car's steering system works. It is also because of the mental image of operating his boat that the same person can predict that by turning the tiller to the left, his boat will turn to the right.
- Mental models help us understand and learn abstract concepts. It is for this reason that the learning of mathematical models is greatly assisted using examples that refer to the personal experiences of pupils / students, compared to the learning process without the use of such examples.

Figure 3.4. Two car dashboards.

We are largely unaware of our constant recourse to mental models. Hints in the environment serve as trigger cues, unconsciously activating relevant models. Predictions are, then, generated based on these, providing shortcuts, and allowing consequent actions to be executed quickly and it is without conscious deliberation that each time a similar set of stimuli is encountered. This feature of mental models is what makes them both economically efficient but also prone to errors.

3.1.3 Eliciting mental models: an example

The elicitation of user mental models is valuable also because it provides us with clues on possible mismatches of user models to the real system, and thus, user proneness to errors. For instance, in one of our studies aiming at redesigning the user interface of a chemical plant with a complex piping system (Figure 3.5), incorporating many valves and reservoirs, plant operators were asked to draw a detailed sketch of the system. The resulting drawings (Figure 3.6) permitted to identify various inconsistencies between the real system and the mental models of the operators. Such inconsistences may prove potentially dangerous when an operator will need to take action in a critical, non-routine, situation.

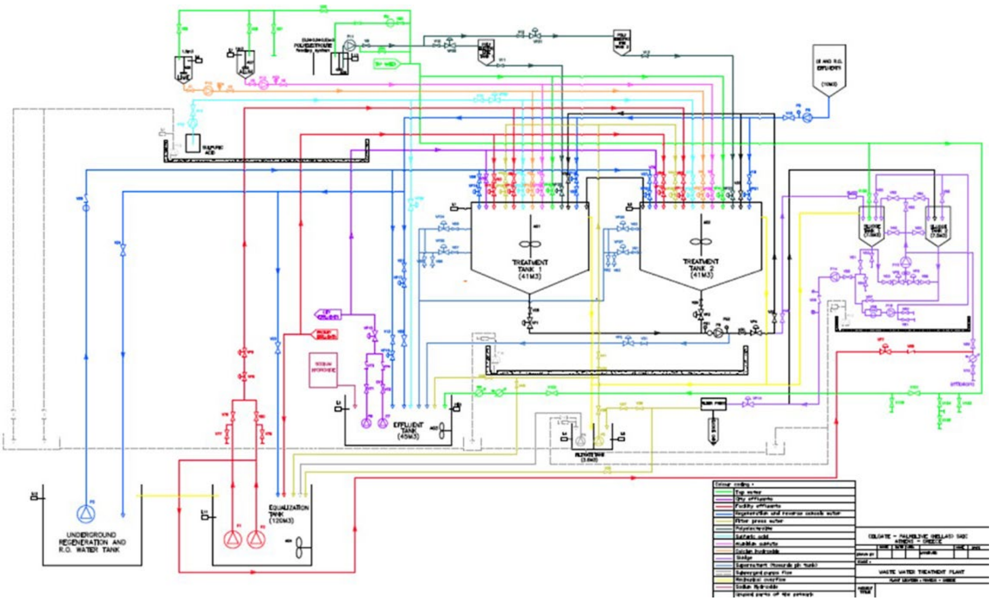


Figure 3.5. The piping diagram of a chemical plant.

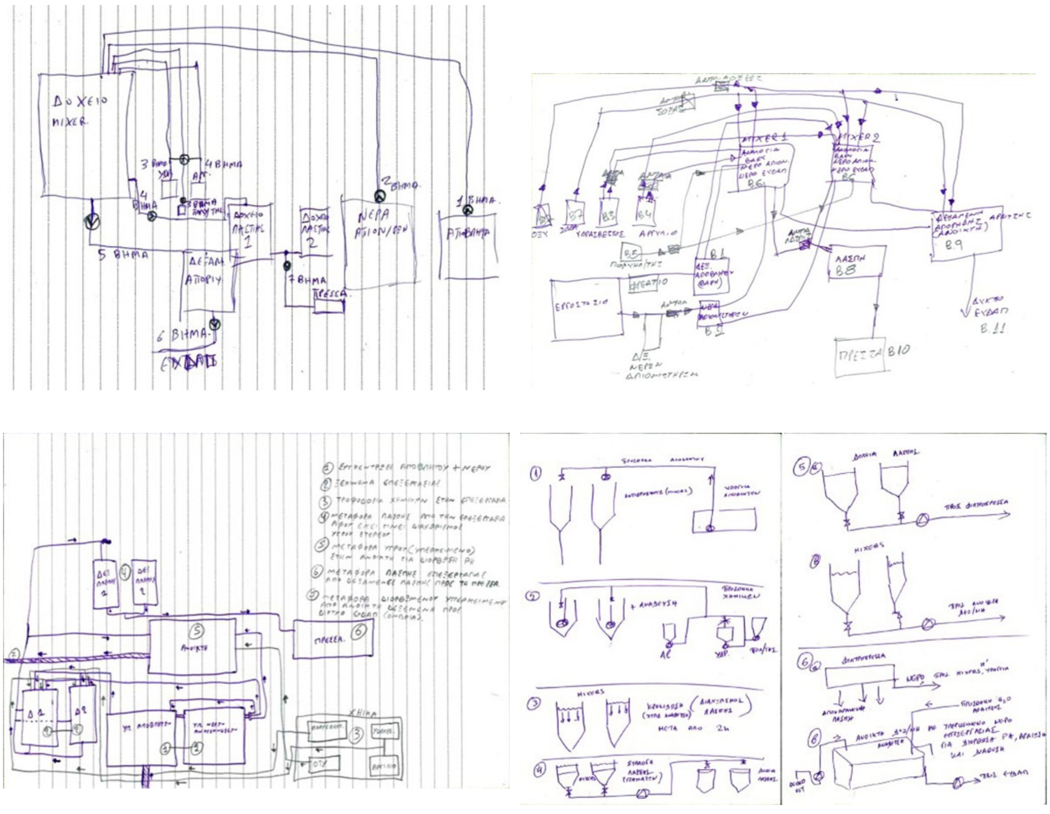


Figure 3.6. Operators' sketches from memory of the piping diagram of Figure 3.5.

Moreover, the sketches revealed quite different renderings of the same system, depending on the operator. This probably reflects different operational schemes (i.e., different aptitudes or styles of control) among the operators. The design of a potential new interface for such a complex system, therefore, should not only be a simplified depiction congruent with the real system but it should also be compatible with the various operational schemes of its operators.

3.2 The Human Action Cycle

In an exercise that we give to our students, the main goal is to observe two novice drivers cooperating to change a car wheel (Figure 3.7). According to their observations, a typical problem that drivers face is the untightening of bolts of the mounted wheel which are generally very tightly screwed. It is commonplace that while they are trying to unscrew them with the bolt wrench by hand in the right direction, they often give up and try the opposite direction. Then again, they change their mind and return to their original plan, but this time by exerting more force to the wrench with their feet. After progressive effort, they finally succeed and continue with this strategy on the remaining bolts. When later asked on the process, they often blame themselves for not being very confident on the unscrewing direction or that better driver training would have avoided such confusion. Truth is that, more than anything else, better design would have easily tackled these issues.

Figure 3.7. Untightening the bolts of a wheel appears to be a demanding task.

3.2.1 The Action Cycle Model

Trying to explain why and when the interaction with artifacts becomes difficult and error prone, Donald Norman (1988, 2013) developed the model of Human Action Cycle (HAC). This concise model examines how people set goals and act in their environment to achieve a desired result. According to the HAC model, achieving the desired result involves two basic cognitive phases: the execution phase and the evaluation phase. It should be noted that, unlike the Information Processing Model (see Chapter 2), which describes cognitive functions, the Human Action Cycle is a model of interaction, i.e., it includes cognitive and physical acts and responses from the environment, without going through the details of internal cognitive processing. Consequently, the two models are complementary; each phase of the HAC model requires the activation of various functions described in the Information Processing model.

The trigger for any human action in the environment is the need to achieve a desired goal (Figure 3.8). In many real-world situations the way to achieve a desired goal is not entirely predetermined – e.g., I want to write a letter. In order to determine the actions that will lead to the desired result, one must decide / choose some ways of action. According to Norman (2013), this activity is called plan setting. For example, to write a letter, I must decide whether to use paper and pencil or a word processor. To implement the plan, the detailed planning of the actions must be done (exactly how and with what succession they will be executed) and then followed by actions in the environment. So, if I decide to use the PC word processor, I have to open the PC, "load" the relevant software, open a new file, etc. Accordingly, to implement these actions I will have to switch on the PC, select the icon of the software with the mouse, etc. This completes the execution process in the environment.

Figure 3.8. The Human Action Cycle (Norman, 2013).

Upon execution of the actions the evaluation process begins. The individual collects information from the environment which are considered to be related to her/his actions, signifies them, checks whether the pre-determined plan unfolds smoothly, and then compares the outcome in relation to the desired result (i.e., whether or not the initial goal is accomplished). So, in the example of writing a letter, I will collect any audible signal emitted by the PC when it is turned on and/or the visual signals emitted by the PC screen (screen illumination, operating system logo, etc.) and signify them as information related to its start-up. Next, I will collect the visual signals associated with the word processor and signify them as information related to loading of the software. In this way, I check that my plan is being executed as planned and that I am approaching the desired result of writing a letter.

According to the HAC model the above processes involve two gaps, named gulfs: the Gulf of Execution, where the individual tries to figure out how something operates, and the Gulf of Evaluation, where she/he tries to figure out what happened after action being taken. The gulfs, large or small are present in all interactive systems and become evident when first interacting with a novel system or device. The role of the designer is to help people bridge the two gulfs. Interestingly, people who experience difficulties often explain them away by blaming themselves. In the case of simple devices —water faucets, refrigerator temperature controls, stove tops—, they expect to be capable of using them, they simply think “I’m being stupid”. Alternatively, for seemingly more complicated devices —sewing machines, washing machines, digital watches, or almost any digital controls—, they may simply give up, deciding that they are unable to understand their functioning. Both attributions of blame are wrong. These are things for everyday household use; none of them have a complex underlying structure. The problem mostly resides in how they communicate their functioning (i.e., their interface design), not in the people attempting to use them.

3.2.2 Consecutive and nested action cycles

It should be noted that the HAC is a generic model of the cognitive processes of carrying out targeted actions. In reality, people do not necessarily go through all steps of the process in every interaction. Very often, when an action is familiar, they take shortcuts, by “automating” parts of the process, due to cognitive skills already acquired. Also, the accomplishment of a complex goal usually requires executing many consecutive or even nested action cycles. In this sense, complex tasks are in turn broken down into sub-tasks and plans. In such complex and non-predetermined processes, there are continuous feedback loops, in which the results of one or more action cycles are used to set the next goals. A typical example is the attempt to use a new device by trial and error. During the execution of the nested action cycles, some of the sub-goals may lead to a dead-end, resulting in the need to carry out new action cycles with completely different sub-goals.

In the aforementioned case of the wheel change process, the participants have a clear goal: to change the wheel of a car. To achieve this, they make a gross plan, i.e., to remove the attached one and then put-on the spare wheel. They specify a sequence of unscrewing the bolts of the old one, lifting the car, removing the wheel, and then, placing the spare wheel, screwing the bolts, and lowering the car to the ground. Each of these steps constitutes a new action cycle. For example, the unscrewing of the bolts needs a certain plan, e.g., “use the bolt wrench”, sequence of actions: “place the wrench on the first bolt, exert pressure counterclockwise, and repeat the same procedure with the remaining bolts”. As explained above, many drivers albeit following this sequence, tend to question their initial correct plan, due the tightly screwed bolts and the lack of feedback, and form a new plan by trying to unscrew clockwise. When they eventually fail again, they revert to their original plan but with more pressure this time, using their feet instead of hands. Any slight movement of the bold (i.e., feedback) will confirm their plan and they will continue with the rest of the routine. So, in this situation there were two distinct problems: the amount of force needed to execute the action and the lack of guidance or confirmation that the counterclockwise direction is indeed the correct one. Similar trivial problems are observed in the process of lifting the car with the jack. We could think of better tools that would signify their proper use (e.g., unscrew by the foot) in the first place, along with a better guidance of bolt untightening direction. Or alternatively we could think of a novel way of changing a wheel with no need to unscrew tight bolts! As we will see in the next chapters, the level of radicalness of our design solution depends on the freedom we have to intervene at higher levels of a system (e.g., car design standards).

3.2.3 The practical significance of the action cycle in design

In practice, the HAC model helps distinguish and clarify how human interaction with a technological artifact can go wrong. Specifically, it distinguishes the intention to act from the ability to act and the perception of feedback from its signification. As we have seen above, it is one thing to understand what you should do and another thing to succeed in physically performing it. Furthermore, the different action stages not only give us insight on how to make an interaction process easier but can also be used in reverse, i.e., to make interaction difficult, so as to exclude certain segments of the population from using a device. Take, for example, detergent bottles and lighters which are dangerous if used by toddlers. Some of the things that distinguish toddlers from older children or adults are their lack of reading skills and their limited physical strength. So, to render a detergent bottle safe for toddlers, we can design an unintuitive two-step cap opening explained through written directions on the bottle. Toddlers will not be able to read them, and thus, form an effective plan on how to open the bottle. Respectively, a lighter can become child-resistant by adding a little guard over the spark wheel or use a high force button (Figure 3.9). This way, although a child can form a correct plan on how to use it, by mimicking an adult, the execution of the plan will be prevented due to the lack of the appropriate physical strength needed.

Figure 3.9. A child-resistant lighter is a lighter that at least 85% of children under the age of 51 months cannot operate due to the lack of sufficient physical strength.

Respectively, in the Evaluation phase the HAC model distinguishes the reception of feedback from its signification, leading to interesting insights in terms of design. For instance, the inability of a user to notice a faint signal whether this is dim light or subtle sound in a noisy environment, require a completely different redesign than a strong signal that can be misinterpreted due to its semantic ambiguity.

According to the HAC model, the ease of learning and using a technological artifact is ensured when the design eliminates or reduces execution and evaluation gaps. The seven stages of the model can be turned into a list of questions that a designer should consider, before implementing specific design solutions (Norman, 2013). These questions are:

How easily will the user of the device / interface be able to:

- understand its function (i.e., what goals can someone achieve with it)?
- understand what actions are possible (i.e., to decide on the ways of achieving these goals)?
- determine the order and manner of carrying out the actions?
- carry out these actions?
- realize that the state of the device has changed since the action?
- interpret / signify the new state of the device?
- conclude if the goal was achieved?

3.3 Design principles for seamless interaction

In the interaction with any technological artifact, the action cycle gaps can be greatly reduced by applying a set of fundamental design principles that facilitate action selection, action execution, perception of feedback and interpretation.

3.3.1 Affordances

Suppose you are given a novel videogame controller that contains no buttons. You already know that this controller is meant to do a particular job, but you cannot figure out how. You turn it around your hand, you press it in various regions, and nothing happens. Eventually, having run-out of ways to do something with it, you start shaking it and aha! The screen opposite suddenly comes to life and seems to obey every tilt of the little plastic piece in your hand.

A similar situation may arise when trying to open a kitchen cabinet or drawer with no observable handle or slider. You stare at it searching for recesses up and down, then, you eventually touch it and the only way to do something with it is to push-it in (a move opposing your intended goal). Nevertheless, after pushing in, suddenly a pop-up mechanism pushes back the cabinet or drawer to open.

In both of the examples above, you have clear intentions of what you want to do but no hints on how to actually execute them. Still, in both cases, you manage to achieve your intended purpose. The reason for this common success is found in the power of "affordances".

Affordances are possibilities for action, i.e., qualities or properties of an object that define its possible manipulations by specific users. The notion was originally introduced by the prominent ecological psychologist James Gibson in 1977, referring to all action possibilities with an object based on users' physical capabilities.

Note that although both "devices" above provide no signs or cues on how to operate them, they subtly indicate their functioning through exclusion i.e., by "allowing" only limited ways to interact with them. In the same way, a handle-less door does not explicitly indicate how it opens; it just affords being pushed, and nothing else (Figure 3.10).

Figure 3.10. The saloon doors offer no cues or signs, they just afford being pushed.

By exploiting the power of affordances, a designer may render an interface legible/discoverable without resorting to any extra signifier (symbolic or other). A successful implementation of affordances in design is the two-finger image zooming gesture for digital touch screens. One could have easily imagined a sliding bar on the bottom of the image or a virtual magnifying glass with (+) and (-) signs to perform the zooming function. What differentiates the two-finger zooming gesture from the above solutions is the absence of such signifiers (i.e., dedicated manipulation elements). The two-finger gesture is not explicitly communicated to the user, instead it is revealed to him by exclusion of other possibilities; it just “affords” to be shrunk or stretched. Once found, the specific gesture feels so intuitive as if occurring naturally (a reason why toddlers try to employ it on printed images too).

Note, also, that affordances may be directly perceivable or not. An opening through a wall or a ramp that affords passing through are both directly perceivable, as is a door with a handle that affords turning. However, a handless door that affords pushing or a digital image that affords stretching, are not directly perceivable; they first need to be explored. Directly perceivable possibilities of action can also be deceptive; for instance, a swell or protrude in a flat surface that suggests to afford pressing (e.g., button) but actually does not, or an underlined text that suggests a hyperlink that cannot actually be clicked.

It is also important to stress that “affordances” cannot be defined without reference to specific user groups. They are not properties or qualities of objects in themselves; they are possibilities of objects offered to specific users with specific abilities. A small flap at the bottom of a door offers an entrance possibility to a cat but not to a human. A narrow ramp offers a motorcycle rider the possibility to overcome a ditch but not to a car driver.

3.3.2 Mappings

Mapping can be generally defined as the relationship between two sets of things. In interacting with a technological artifact, we may distinguish between Control – Effect mappings and Effect – Feedback mappings.

Control-Effect mappings are correspondence relations between controllers or control actions and their target effects on system function. The most pertinent example of mapping is a directional / topological one. In a joystick that can tilt in all directions one expects the controlled target to move to the same direction as the joystick. In a ship with two propellers, one left and the other right, one expects the control throttles to be placed correspondingly. Such mappings, although seemingly trivial, allow users to intuitively form action plans (i.e., match controls to outcomes), without any need for explicit signifiers.

Mappings can also be functional. A controller that has X discrete positions corresponds well to a target effect that also has X discrete states (e.g., automobile gears), or a controller that has continuous feed corresponds well to a continuous variable target effect (e.g., the automobile accelerator pedal). Also, progressive force feedback control corresponds well to exponential target effect (e.g., automobile brake pedal).

Mappings can also be semantic in nature. Shape, color, or other qualities of controllers can be designed so as to match the target effect. A two-handle kitchen faucet can be directly used if each handle is colored correspondingly with red (for hot) and blue (for cold). Two switches, one small and one large, can easily be corresponded to two electrical circuits, one low voltage and one high voltage. Semantic mappings are typically used in multi-control centers where many action controllers are grouped together according to their meaning for the controlled system (e.g., all landing gear controllers grouped together in an airplane cockpit).

Effect-Feedback mappings are correspondence relations between feedback signals and their target effects on system. They are important for narrowing the evaluation gap in the action cycle. Good correspondence of feedback signals with the target effects (be them topological, functional, or semantic) help in correctly interpreting these signals. For example, a "BUZZZ" sound is well semantically mapped with a failed action, while a "CLING" sound with a successful one.

3.3.3 Behavior-shaping constraints

Behavior-shaping constraints (hereafter constraints) are strictly speaking restrictions of possible user actions. Their utility in the design of interaction can be important for two reasons: (i) they can safeguard against unwanted or hazardous actions, and (ii) by restricting certain actions, they indirectly point towards the correct way of performing an action or sequence of actions. The most effective constraints are physical ones. Physical constraints are the flip side of affordances; they literally prohibit (or restrain) an action from taking place in conditions. For example, the shape of an electrical plug not only affords to be connected to the appropriate socket but also physically constraints it to be connected to certain sockets (e.g., 110V / 220 V). In the same manner, a fire door affords being opened from the inside of a building, but on the flip side, it physically constrains someone from opening it from the outside. Another example is shown in Figure 3.11.

Figure 3.11. The cut-out of the SIM card is purposefully designed as a physical constrain to prevent erroneous placement of the card in an electronic device.

Physical constrains to prevent human errors have been first formally introduced in industrial processes in the 1960s under the term "Poka-yoke" i.e., "mistake-proofing", as part of the Toyota Production System.

Behavior-shaping constraints can also be cognitive. Cognitive constraints do not physically prohibit users from performing a certain action, but merely inform them of what not to do (Figure 3.12). Cognitive constraints are implemented through various signifiers.

Figure 3.12. Error proofing Medical Gas outlets in hospital incorporating both physical and cognitive constraints.

3.3.4 Signifiers

As introduced in Chapter 2, signifiers are the body of signs, i.e., perceivable indicators that convey some information, termed "signified", to a user. Thus, for instance, written labels are "signifiers" that convey the meaning of the written text, the "signified". Signifiers may be designed, as symbolic or iconic signifiers (e.g., a STOP label or a BUZZ sound) or naturally occurring as in indexical ones (e.g., a recess in a drawer or an engine noise). Signifiers play an important role in helping users to specify action plans and translate feedback signals to meaningful interpretations.

Note that signifiers are often embedded in affordances and mappings. A perceivable affordance acts as a natural signifier of itself (e.g., a recess in a drawer), as it is naturally linked to its signified meaning, whereas a semantic mapping may use signifiers to achieve correspondence between control actions and effects through socio/cultural conventions (e.g., red and blue coloring in faucet handles).

3.3.5 A case in point: a ventilation interface

In our laboratory there is a simple wall ventilator with an interface as depicted in Figure 3.13. Although it is a quite simple device, most new users find it difficult to understand. The first trivial challenge users face is the direction in which they have to move the upper slider to switch-on the system. When asked before touching it for the first time, some people suggest it should be slid right while some say left. This ambiguity is caused because both the outer frame and the inside rectangular "button" with the embedded signifiers (i.e., 0/I) could be perceived to afford sliding. Although unintuitive to some, the correct action is to slide the outer frame to the right.

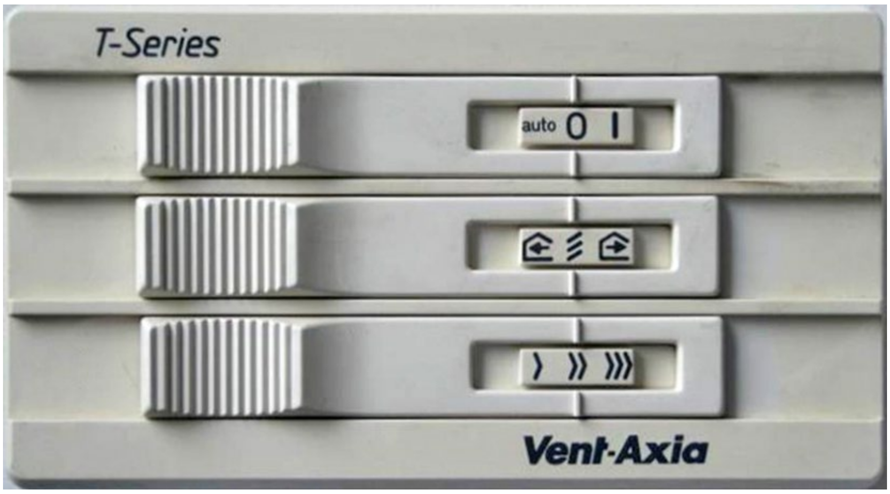


Figure 3.13 The interface of a wall ventilator.

The next challenge faced by novice users is to interpret the three states of the second slider. The two outer states of the slider are quite obvious, thanks to their respective signifiers; they change the fan's rotation direction between sending air in or out of the room. However, what about the middle state with the three inclined lines? In this state, the ventilator fan is stationary, but with the shutters open, allowing air to passively flow through them. Only few people will make a correct guess on this. The last slider is rather straight forward; it controls the fan speed with three discrete levels.

Besides the two points above, what makes this interface difficult to grasp on the whole, is that the novice user cannot form a clear mental model of its various modes. This is primarily due to the grouping of its controls that reflects the ventilator's electromechanical functioning rather than its intended use modes (i.e., a technology-centered design as opposed to a user-centered one).

Technically, the specific ventilator is a single-phase AC motor, its circuitry functioning as follows: the power supply is controlled by the top slider; once this slider closes the circuit, a simple linear actuator opens the shutters, and the current reaches the starter winding switch. Moving the second slider either right or left, closes the starter winding circuit, activating the motor clockwise or anticlockwise depending on the chosen polarity. The last slider controls the rotation speed through activating the appropriate main winding (one for each rotation speed). Note that ensuring that the shutters are open in any of these states is crucial to prevent the motor from overheating. Thus, connecting the shutters opening with the main power switch makes perfect sense engineering-wise. The open shutters, even without fan, provide a useful function: they allow the air to flow through the wall opening. However, this is not communicated to the user apart from a rather obscure "shutters open, motor Idle" indicator in the middle position of the second slider. Many users, therefore, fail to grasp the "shutters open, motor Idle" function as its indicator is located in a control dedicated –seemingly and physically– to the fan's direction of rotation.

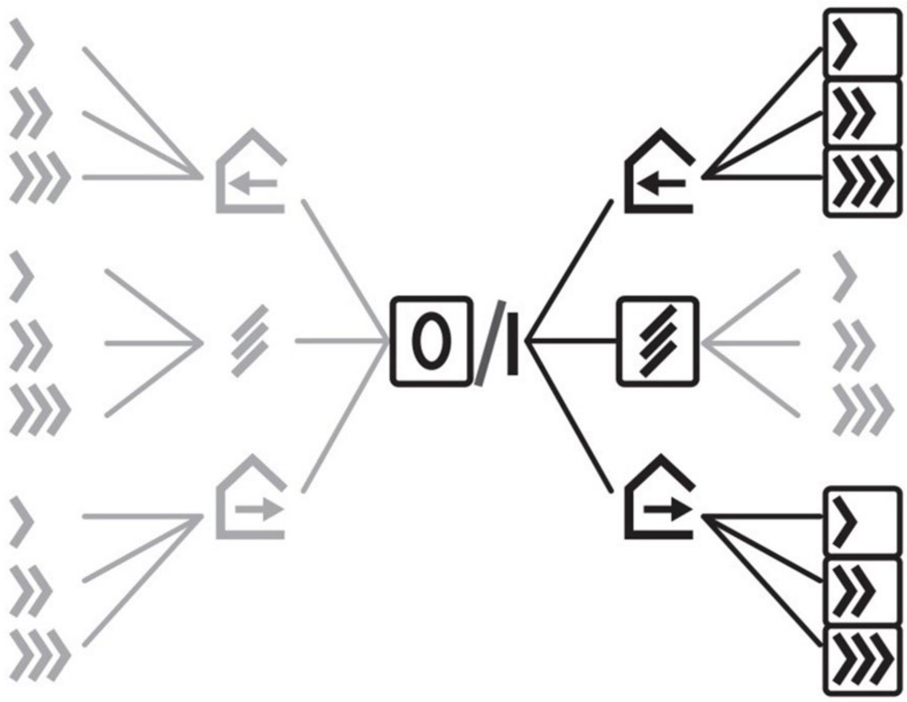


Figure 3.14. All 18 possible configurations that the ventilation interface allows. The eight functionally meaningful states are framed while the meaningless ones are greyed out.

A further issue with this interface is that it allows the user to set many unnecessary or meaningless configurations, in other words, there is a mismatch of “offerings” and actual meaningful states. The interface affords in total eighteen (18) different configurations (three independent sliders, one with two, and the others with three potential positions i.e., 2x3x3) (Figure 3.14). Most of them are meaningless –like having set the speed slider to maximum while the middle slider is set to passive flow, or the upper slider is set to “off”. There are only eight meaningful states of the system which are presented in Figure 3.14. (Three fan speeds for each of the two rotation directions, plus a passive flow of air through open shutters and the zero state where the motor and shutters are off).

One of the principles of intuitive interface design is to match all possible configurations of the interface to the distinct functional states of the system. For example, the alternative design of the ventilator interface shown in Figure 3.15, communicates better all the functional states of the system, as it affords only the eight meaningful configurations, leaving no space for misinterpretation even before interacting with it. There is still a downside in this design; it lacks “memory”. A frequent user must reset the system to the desired state. While on the previous design they could simply switch it on and off leaving the other sliders on their “favorite” setting. As no design solution is perfect, the designer must each time be aware of the advantages and disadvantages of each solution and to whom they should prioritize (e.g., first time or frequent user).

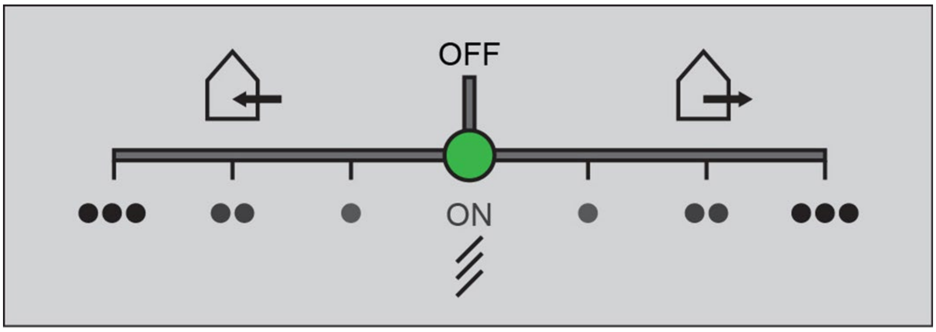


Figure 3.15. Alternative interface design concept for the wall ventilator.

3.3.6 Interface configurations and meaningful states

In the ventilator example we have come across, there is a mismatch between the configurations that the interface affords and its meaningful functional states. In most cases, such mismatches are solved through proper signifiers on the controls (e.g., an “on/off” power button would imply that when set to “off” all other controls are disabled). However, such obvious associations apart, there are many more that a user would be unaware of before interacting with the system. The common argument for accepting such meaningless configurations in a system is that they are harmless (i.e., since not associated with a real function, nothing will happen). However, as evidenced by numerous accidents in safety-critical systems, even the slightest ambiguity caused by arbitrary control associations, may drive users to form flawed mental models which in turn undermine their control and diagnostic abilities.

Undesirable interface configurations can be divided into three categories:

- Harmful to the system (e.g., closed shutters + fan operation in the ventilator example).
- Technically meaningless (e.g., system off + maximum fan speed).
- Technically feasible but not covering any useful scenario (e.g., setting exact RPM speed on the fan or seconds on an ordinary alarm clock as in Figure 3.16).

Figure 3.16. Offering second level precision when setting a wake-up alarm comes with a cost: added complexity and more room for error, when setting the alarm time, just to cover an admittedly marginal scenario.

To minimize possibilities for error, configurations harmful to the system should be eliminated; technically meaningless and useless configurations should be best avoided to minimize clutter and ambiguity. Especially in safety-critical systems, user understanding of the interface should not rely only on the interpretation of signifiers. The interface should communicate acceptable system functions, as far as possible, through its perceived affordances, before the potential user actively starts interacting with it.

3.4 The Skill-Rule-Knowledge model of cognitive activity

As noted above, the Human Action Cycle is a rather simplified representation of the cognitive process taking place during human interaction with technological artifacts but does not include or explain the shortcuts an experienced user takes when performing a familiar task. In Chapter 2 we have seen how humans can simultaneously perform multiple tasks when some of them are familiar and do not require their full attention. Indeed, many or our daily actions are more or less "automatized" requiring only brief conscious action cycles. Take, for example, the case of driving. A novice driver will be much focused on the task of controlling his vehicle, being fully attentive when changing gears or steering, passing through all stages of the action cycle: continuously making new plans, acting and anticipating feedback from the car and surrounding road environment. An experienced driver though, will not consciously focus on such tasks but rather on driving strategies like how to take the fastest route or drive more fuel efficiently. The purely procedural and sensory-motor parts of driving will be performed without any meticulous planning or any longing for feedback.

To confront this impediment, Rasmussen (1983) proposed a classification of cognitive activity in human machine systems. According to this classification and the accompanied heuristic model (Figure 3.17), we may distinguish three general levels of cognitive processing: Skill-based, Rule-based, and Knowledge-based (hence its name SRK model). When confronted with a novel situation in a specific domain, we employ general knowledge related to the structure and operation of the domain (i.e., our mental model of the domain), to formulate assumptions on which to base our actions. Actions taken will eventually be checked by trial and error, to allow us to reach the desired result. This level of cognitive process is called knowledge-based and is actually identical to performing successive full actions cycles.

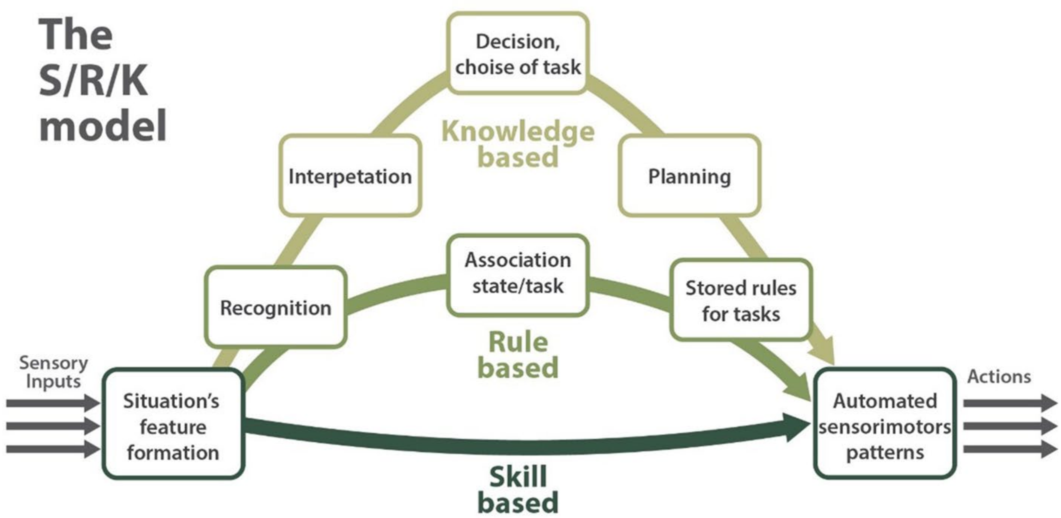


Figure 3.17. The Skills–Rules–Knowledge Model of human activity.

After some recurrent action cycles, we tend to generate rules that help us ease our cognitive processing (e.g., formulating assumptions). These rules can either be created to be used only by ourselves or be taught to others so as to help them become more efficient sooner. In the first case, if at any given time the rules do not provide satisfactory results or cannot be applied for some reason, we can always revert to the knowledge-based level and conduct a new effort to resolve the situation. Instead, if the rules were given to somebody (e.g., in the form of a procedure), the eventual lack of knowledge concerning the structure of the work domain will prevent him from working at a knowledge level and the situation might be unsolvable.

Becoming more and more accustomed to a habitual procedure through practice, we may start to gradually internalize it, to the point that it becomes embodied, as we saw in Chapter 2. Then, we act based on automated sensory-motor schemata, while most of our conscious attention can be focused on other tasks. This is the skill-based level of activity, only disrupted when an event suspends the familiar flow of the process and redirects our attention to it. Then we are called to consciously deal with it either on the rule-based level or the knowledge-based level if possible. So, while working on rule level requires the use of declarative memory, the skill level is based on the procedural memory which as we saw in Chapter 2, is structurally embedded in the mind and body of the person without the need of conscious recollection. An important consequence of the above is that in critical situations where many cognitive resources are either allocated elsewhere or even blocked (due to anxiety or confusion), the skill-based actions can largely continue unhindered.

Back to the driving example, an experienced driver that commutes every day, following the same route, he/she does so, at a skill-based level without allocating any cognitive resources on planning or questioning the route. If, however, one day the usual route is blocked, the driver will shift his/her mental focus to resolve the issue based on rules built in the past. S/he may, thus, take an alternative route s/he had used some time ago and that would resolve the problem. If by chance the alternative route is also blocked, then s/he must find a solution based on his/her orientation skills and his/her geographical knowledge of the area and road network. If the route taken proves to be successful, a new rule might emerge that will be considered next time it is needed. However, in the unfortunate case that the driver is new to the area and has relied on given rules (or follow road signs) to get to his/her destination, s/he will not be able to cope with the situation due to his/her inability to revert to the knowledge level (i.e., to the lack of an adequate mental model of the specific road network).

Interestingly this is often the case in complex industrial systems, when the operators are trained predominantly at a rule-based level. At this level of cognitive activity, the action cycle is hindered in terms of planning, and system feedback is interpreted based solely on established rules. The operators' inability to revert to the knowledge-based lever, due to fragmentary or deficient mental models of the underlying physical process, has impeded them in averting many industrial disasters in the past.

3.5 Errors and human activity models

It might seem strange at first, but there is no unequivocal scientific definition of human error. Humans may fail to accomplish a task due to various factors. However, to designate a failure as an error requires not only a formal causal link between the failure and specific human behaviors, but also a judgement of such behaviors as inappropriate from a specific point of view. Therefore, errors are primarily “attributions” of cause, not facts. Nevertheless, an authoritative attempt to defining human error is provided by James Reason (1990) who defines it as “a generic term to encompass all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency”. Note that according to the above definition, human errors are attributions rather than facts, and can only be characterized as such after the fact, i.e., after an unwanted outcome occurs and human actions considered to have contributed to it. Thus, error attributions may differ depending on who is the attributing agent (e.g., the person performing the task or an external observer) and on the space/time horizon considered for the outcome (e.g., failure to set the intended temperature in a thermostat –a momentary simple task–, or failure to correctly diagnose a disease –a complex and laborious cognitive process).

Putting these theoretical subtleties aside, the SRK model helps us to identify the level of cognitive processing that such deficiencies may occur (Figure 3.18). If the deficiency is at the level of specifying or planning of activities, then it is qualified as a mistake. According to the SRK model mistakes can happen at two levels, at the knowledge-based level, i.e., defective generation of plan of action, and at the rule-based level, i.e., defective selection of a plan (or rule). In the first case, a defective plan may be attributed to an inappropriate mental model of the situation or a decision in the face of uncertainty that resulted in goal failure, while a defective selection of rule may be attributed to misinterpretation of the situation at hand caused by a complex rule structure or ambiguous signs. If the plan of action was adequate but failed at the level of actual performance, (i.e., a mismatch between plan and its execution) then it is qualified as an action slip or memory lapse. Action slips and memory lapses typically happen at the skill-based level and are thus unintended, in the sense that are caused by the activation of a wrong routine. Slips and lapses may be attributed to many situational and person specific factors, however proper physical interface design can tackle many of them.

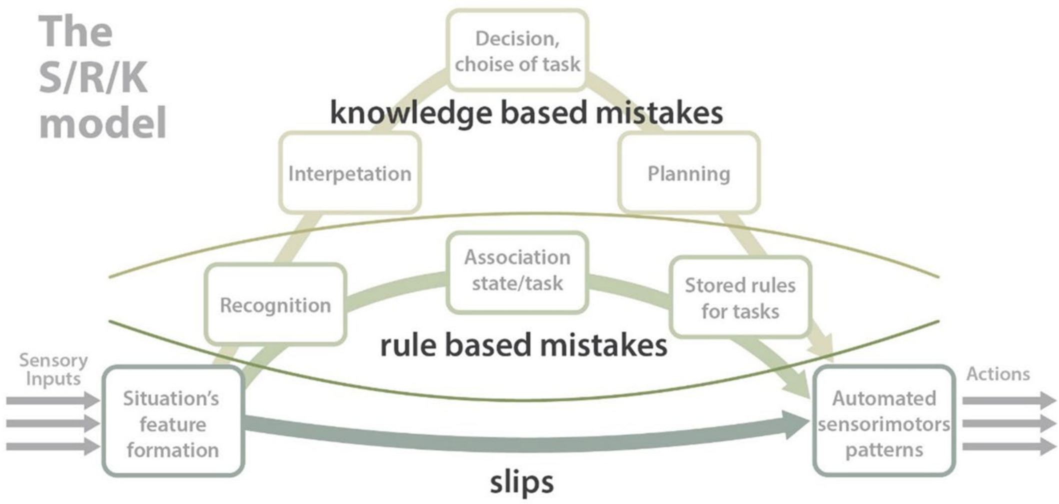


Figure 3.18 The attribution of slips and mistakes in the SRK model.

The above error types can also be mapped in Norman's action cycle (Figure 3.19). Mistakes map directly at upper levels of the execution phase of the action cycle. Examples of such mistakes are wrong goal formation, inadequate intention specification, or poor action plan. Slips and lapses typically occur at the level of the execution phase. For instance, the unintended pressing of the wrong button or the unintended omission of one step in a habitual action sequence (see also example in Figure 3.20).

Persistent slips in a human-machine interaction can be rectified by various design attributes such as control differentiation, providing confirmation steps or proper interlocks in critical parts of human input. However, such design elements are unlikely to prevent from mistakes (i.e., intentional errors). Mistakes are trickier to tackle but they can be minimized by proper interface design, which facilitates the formation of appropriate mental models.

Errors at the evaluation phase of the action cycle can be classified according to SRK levels. An omission in perceiving a feedback signal can be attributed to the skill-based level of behavior, while errors in perception and interpretation of feedback can be attributed to the rule- and knowledge-based levels depending on the nature of the task.

Persistent omissions in signal perception can be rectified by strengthening relocating or changing the modality of the signal(s). Persistent interpretation and/or evaluation errors can be mitigated by redesigning signifiers and mappings on the interface.

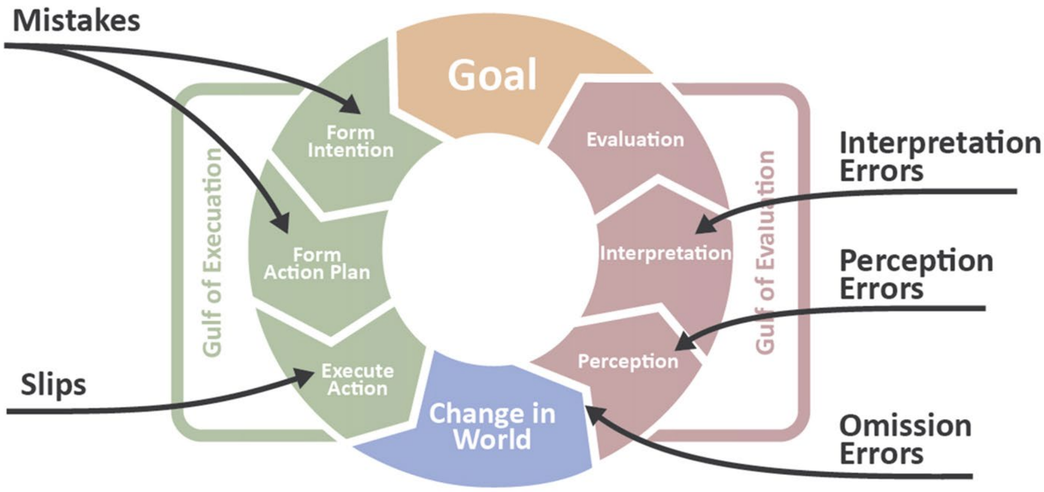


Figure 3.19 The different kinds of human errors mapped on the Action Cycle.

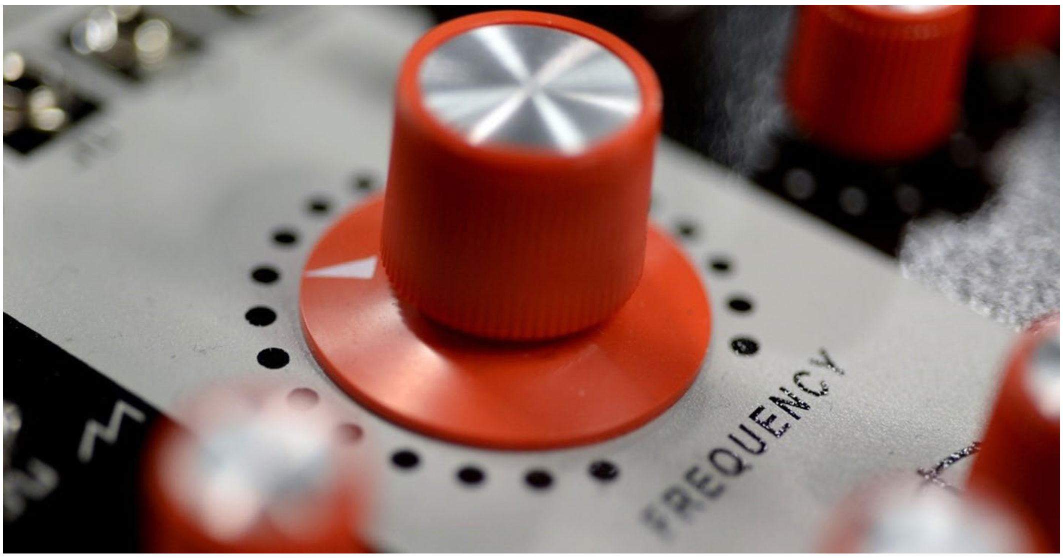


Figure 3.20. Failing to set this knob to the exact desired value, due to the tight configuration of possible states, is considered a slip. In this case, the intended action is probably correct, but the user fails in terms of the dexterity needed. Instead, if the user wanted to adjust the volume but decided to turn the frequency knob instead, then, this is considered as a mistake. At the same time, if the knob provides no tactile feedback when switching between states, this may lead to omission errors.

Last, it is important to stress that there is no simple dichotomy between errors and correct behavior: from a design perspective, observed errors indicate a mismatch between user abilities and system features. Therefore, the entire interaction should be treated as a “cooperative endeavor” between person and machine, one in which misconceptions can arise on either side (Norman, 1988).

3.6 Complex cognitive tasks

During the presentation of the SRK model, we saw that when we encounter a novel situation, we enter into a knowledge-based cognitive process, conducting more or less consecutive complete action cycles. In this section we will focus on sources of complexity and on ways to handle them at the knowledge-based level.

These cognitive processes can range from relatively simple to quite complex, effecting considerable mental load. The mental load depends on the complexity of the system being managed, as well as on the cognitive abilities of those who are called upon to manage it. For this reason, in cases of complex systems management, the term complex cognitive task is used in the international literature. Typical examples of complex cognitive tasks are the planning and control of air, land or sea traffic, the operation of power plants and electricity distribution networks, medical diagnosis, stock-market decision making, as well as systems design, as we will see in Chapter 4.

A work-system is considered as “complex” when we identify one or more of the following characteristics:

- it consists of many interdependent-interacting elements or parameters,
- is dynamic, i.e., the elements that comprise it evolve over time and the parameters that describe their operation take different values over time,
- events that affect its operation can occur at different time frames (these events can come either from the system itself or from its environment),
- there is uncertainty as to when these events may occur, both in their nature, and as of the consequences they may have on the operation of the work system,
- there are many, prioritized or not, quantitative and qualitative goals, which the operation of the work system should achieve, and which can often be at least partially conflicting (e.g., quantity, quality and production speed, safety, resource savings),
- it imposes strict time limits on human operators,
- there is poor reliability of the work system and/or human erroneous actions entail significant risk.

Despite the incessant progress being made in the direction of automation, human contribution continues to be necessary, not only for the design of systems, but also for their control and diagnosis/troubleshooting of possible malfunctions, for instance in modern aircrafts or in process plants. Despite the increasing number of intelligent control systems with which such systems are equipped[6], the presence of specialized operators is still –and may always be necessary at some level to control– for unanticipated events that go beyond automation capabilities. These operators are, therefore, confronting less frequent but increasingly complex cognitive tasks.

3.6.1 Coping with complexity

To illustrate how easily a process may become complex and to the difficulties that arise in terms of monitoring and diagnosing due to this complexity, we will present an example of a hypothetical –rather simple– experimental set-up.

Consider a biological experiment that takes place inside a sealed chamber (Figure 3.21). A light bulb that is of outmost importance for the purposes of the experiment should be always lit. There is no physical opening or any visible access at the internal space of the chamber. The switch of the lamp is outside of the chamber and our sole concern is to be able to check that the light in ON. Since we have no direct visible access to the interior of the chamber, we would want to make a pretty reliable system to monitor the lamp. Our first concern would be to ensure that the switch outside of the chamber that closes the circuit is visibly ON. To do this we would probably choose a two-state switch that clearly communicates its current state (Figure 3.21a). However, the switch does not warrant that the electrical power will reach the bulb in case of wiring failure. So, a second control point would probably be used to incorporate a second circuit connected to the lamp holder that gives a light signal outside the chamber, whenever electrical power is present (Figure 3.21b). Now we are informed that electrical power reaches the interior light bulb, but is the second circuit more reliable than the first one? If the external light (no2) is OFF, are we certain that the interior light is OFF too? There might be a failure in the second circuit or the signaling light. To further reassure that the signal light is functioning we should have a small test circuit that can be used to check it (Figure 3.21b). Indeed, if the light no2 is functioning we can be certain that the internal lightbulb receives electrical power but not that it is actually lit. To ensure that the light bulb is actually lit we should install another monitoring device such as a light meter (Figure 3.21c). Such a device would give us very accurate indications of the actual phenomenon that really matters, but its complexity makes it more prone to failure. So, if the switches and test lights indicate that everything is functioning fine, but the light meter indicates the opposite, then we wouldn't be very confident to diagnose the possible failure. That would perhaps lead us to install yet another measurement device, such as a thermometer to detect radiated heat from the lightbulb (Figure 3.21d). Of course, a thermometer is also a device with less than 100% reliability, and we could not count exclusively on this but rather on the combination of all the above control points.

[6] i.e., decision support systems such as rule-based expert systems, neural network/machine learning algorithms or various operation research methods for optimizing decisions.

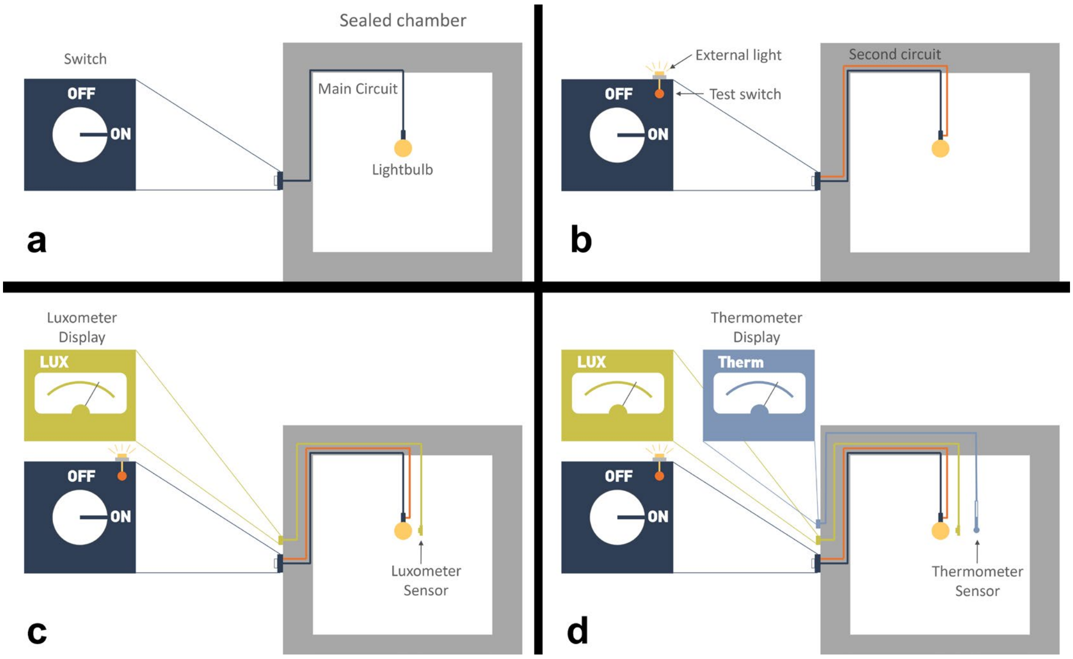


Figure 3.21 A biological experiment set-up. As no direct visual inspection is possible inside the chamber, alternative sensors are installed to monitor that the yellow lamp remains lit. The presence of multiple sensors renders the system more reliable but also more complex to diagnose in case of incompatible sensor readings.

Undeniably our system has become more reliable on the whole, but what about our ability to diagnose possible inconsistencies between the sensors or potential failures? What if the light meter shows proper measurements and the thermometer showing lower temperature than expected? It might be that the light bulb has just been lit and the temperature has not yet risen or that the thermometer is malfunctioning or that the light is off and the light meter malfunctions. All these potential scenarios have their own probability based on the reliability of each sensor, but the system supervisor is unable to compare the odds of all the potential scenarios, especially under pressure to make a critical decision. Such is the challenge with complex system monitoring like nuclear power plants or modern aircrafts where the pilots have no direct access on the mechanical parts of the plane and must solely depend on the numerous sensors that provide information about the airplane's condition.

A typical method often employed to rectify the above problem is to combine the readings of different sensors through an "expert system" (i.e., a set of logical and/or probabilistic rules) that presents ready-made or computed logical inferences set by system experts.

Let's have a look at a simplified example. Suppose we have a closed water pumping system that consists of the following parts: a motor, a belt-drive and a water pump with sensors installed at each part, as shown in Figure 3.22. To overcome the diagnosing difficulty faced in our previous example, a rudimentary expert system is installed, presenting written messages resulting from predefined state combinations.

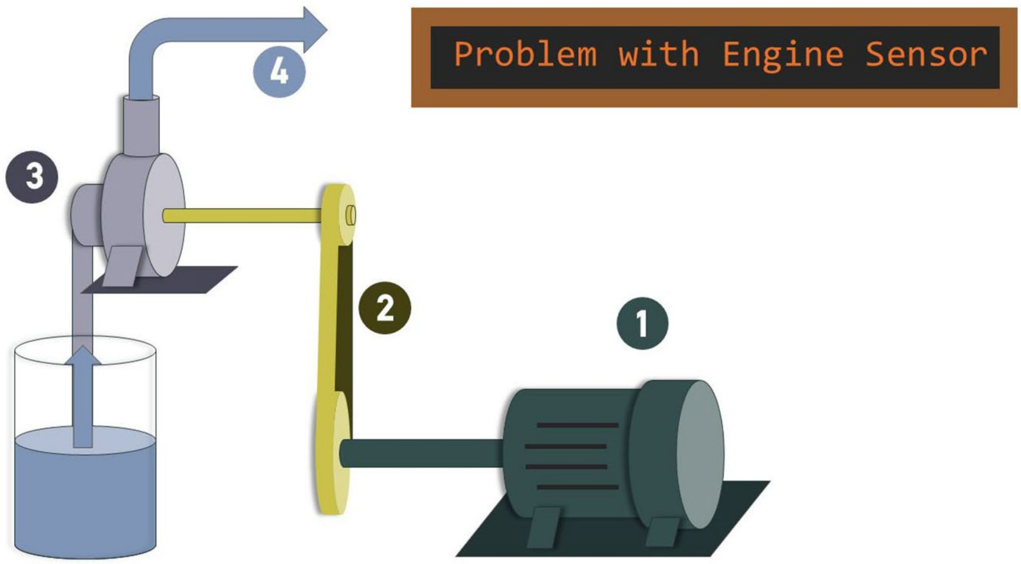


Figure 3.22. A water pumping system consisting of: a motor (green), a belt-drive (yellow) and a water pump (grey) monitored through a system consisting of sensors (1,2,3,4) and predefined logical inference combinations (upper right).

For example: If "1" is OFF and "2" AND "3" AND "4" are ON, then a message "PROBLEM WITH ENGINE SENSOR" would appear because the operation of subsequent parts depends on the operation of the motor, rendering rather improbable that the motor is OFF and all subsequent sensors faulty. But could the expert system algorithm present the same message under the combination: if: "1" is OFF, "2" and "3" are ON and "4" is OFF? Supposing that all sensors have the same level of reliability, could we take the risk to say that if two of the sensors are giving positive signals, then it is likely that the first and last sensors are malfunctioning, or the opposite? Many such ambiguous situations may arise depending on the possible combinations of sensor states. In reality, such decisions are taken through probabilistic risk analysis and various heuristics, sometimes promoting system safety and sometimes productivity. In any case, having all these sensors and having made an a-priori analysis of all possible states, setting rules and guiding instructions to the end user of the system, saves valuable time in decision making and increases the overall reliability of the system. The downside of this approach is that the end user, being blind of the underlying system structure, cannot form an adequate mental model of the system and is thus restricted to work at a Rule-based level. This can result in inability to recognize and intervene when a non-predicted situation or malfunction of the expert system occurs; often leading to cascading failures.

3.6.2 Cognitive biases in diagnosis

In Chapter 2 "cognitive biases" were introduced, as specific human cognitive peculiarities which affect us when interpreting events that occur around us. At this point, it is appropriate to refer to certain cognitive biases of experienced workers that may lead to incorrect or non-optimal diagnostic practices. It should be noted that cognitive biases are in a sense the "negative side" of users' cognitive skills and can be attributed to the latter's attempt to simplify the mental process of diagnosis; in other words, to reduce the cognitive load it entails, as well as succeed in addressing system malfunctions in the shortest possible time. Pioneers in the study of cognitive biases are Tversky & Kahneman (1974), who studied decision-making in uncertain environments. Along with the presentation of cognitive biases, we will consider some possible ways to eliminate, either the biases themselves or their negative consequences. Of course, these methods should always be combined with the appropriate worker training, which remains pivotal for improving diagnosis and decision-making performance.

- Users often tend to ignore information they find unlikely. Thus, for example, there is a risk that they attribute this information to a malfunction of the control systems and not to a malfunction of the actual system itself. To reduce the likelihood of negative consequences of this bias, the design of systems should be such as to enable users to check the reliability of the information they receive or, if possible, to allow them direct access to the actual process (e.g., providing portholes to view parts of the controlled system). Another measure, applied in various domains, is the existence of continuous indication of control systems' functioning and sensors status.
- Users often avoid consulting sources of information they have difficulty either to access (e.g., being located far from them) or to read (e.g., long text or in very small or faint characters). Furthermore, they avoid considering information that is difficult to understand, even if these difficulties are offset by the importance or usefulness of the information they transmit. The most effective measure to avoid this cognitive bias is the ergonomic design of the mediators.
- When searching for a malfunction in a system, operators often tend to stop gathering information as soon as the malfunction is linked with a similar situation that has been successfully dealt with in the past. They, thus, quickly attribute to the present malfunction the same causes as the past one and take the same actions to deal with it. This bias results in the omission and/or non-search for additional available information that may disprove their current diagnosis. This bias can be reduced trough proper interface design, so as to provide all necessary information in a format that reflects the exact state of the system at any moment.
- When generating assumptions about the causes of malfunctions, significant discrepancies are often observed between the subjective and the actual probabilities of the causes. So, for instance, causes that have been identified for a malfunction that has occurred in the recent past, are attributed much higher chances than the actual ones. This cognitive bias result

in spending valuable time and effort investigating these causes that are mistakenly considered more probable and thus delaying diagnosis. Here, the improvement of users' knowledge through presenting historical-statistical data on occurrence frequency of causes of past malfunctions can prove quite useful.

- Finally, another typical cognitive bias in diagnosing is that, once the operators come up with an initial hypothesis about the causes of the malfunction, they tend anchor on it, even if subsequent information reaching them in the meantime makes some other hypothesis more likely. Here, teamwork and training the users against this bias can reduce the negative consequences of such a situation (see for example Crew Resource Management training in aviation).

Illustrative of the above is the following accident.

Just before midnight on December 29, 1972, a Lockheed L-1011-1 TriStar, Eastern Air Lines Flight 401 from New York JFK to Miami crashed near the landing airport causing 101 fatalities. The crash occurred while the entire cockpit crew was preoccupied with a non-functioning landing gear indicator light, failing to notice that the autopilot had inadvertently been disconnected, and as a result, the aircraft gradually lost altitude and crashed.

The flight was routine until 23:32, when the plane began its approach into Miami International Airport. After lowering the gear, the First Officer noticed that the landing gear indicator had not been illuminated. This was later discovered to be due to a burned-out light bulb. The landing gear could have been manually lowered, nonetheless. The pilots inspected the landing gear, but still failed to get the confirmation light (Figure 3.23).

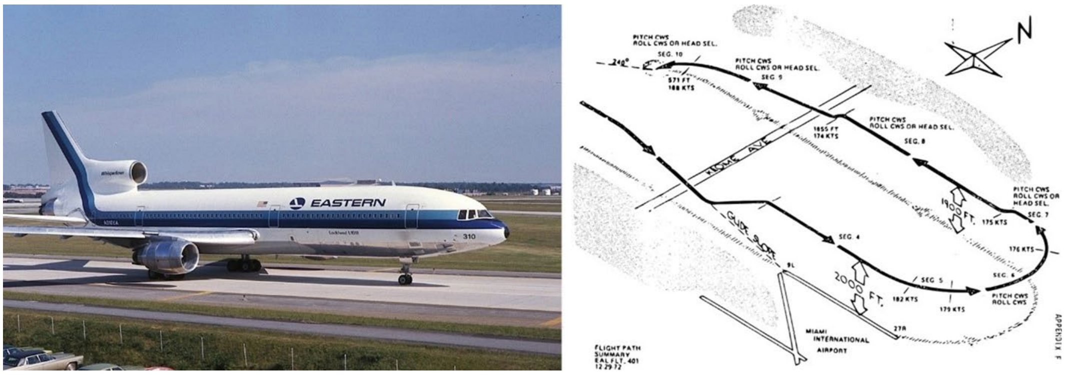


Figure 3.23. N310EA, the aircraft involved in the accident and its flightpath summary, as shown in the NTSB report.

The final NTSB (National Transportation Safety Board) report cited the cause of the crash as pilot error, specifically: "the failure of the flight crew to monitor the flight instruments during the final four minutes of flight, and to detect an unexpected descent soon enough to prevent impact with the ground. Preoccupation with a malfunction of the nose landing gear position indicating system distracted the crew's attention from the instruments and allowed the descent to go unnoticed".

3.6.3 Troubleshooting malfunctions in complex systems

In complex systems, diagnosing malfunctions and restoring them is rarely a linear stepwise process. Indeed, given the nature of these systems, it is not always possible to separate the two phases in time, i.e., first to identify the causes of the malfunctions (diagnosis) and then, to determine how to restore them (solution). This is often due to evolving phenomena (e.g., in an ongoing chemical reaction or in a ship sailing in constrained waters) that require from operators to take immediate remedial action against imminent dangers, e.g., to isolate critical parts of the system, or to render the system stable, before the exact causes of the malfunction are identified. In other cases, if the causes of the malfunctions cannot be identified after an initial diagnosis, operators take actions based on hypothetical causes, to eliminate them one by one[7]. In this “diagnosis by trial” strategy, an action may either restore system functioning, or may provide new information for further hypothesis formation. Therefore, remedial actions also serve for diagnostic purposes, until the exact causes of the malfunctions are identified. All the above nonlinear think-act sequences for rectifying system malfunctions are categorized under the term “troubleshooting”.

The troubleshooting strategy is often unavoidable when facing malfunctions with more than one cause. In such multi-cause malfunctions, it is extremely difficult to form solid initial hypotheses. This is due both to limits in operator cognitive processing, but also because, in such complex events, system algorithms may not be able to cope, thus providing misleading information.

3.6.4 The Decision Ladder: mapping cognitive activity in troubleshooting

The Decision Ladder is a model for mapping operators’ cognitive activity during troubleshooting in complex systems based on the SRK classification of cognitive activity. It was developed by Rasmussen (1986, 1994), following a series of analyzes of real-world malfunctions in high-risk technology systems. According to the Decision Ladder, a troubleshooting process begins with the perception of some information, which acts as indication that a malfunction is occurring in the observed system (Figure 3.24). The perception of this information puts the operator on alert and so s/he begins to collect additional data on the state of the system. Based on the collected data, as well as from the recall of relevant mental models, the operator creates an image of the current system state. S/he, then, interprets the consequences this situation may have

[7] Diagnosing by trial is also quite common in medical practice, termed differential diagnosis. In differential diagnosis, a doctor differentiates between two or more conditions that could be behind a patient’s symptoms, and progressively provides trial treatments to eliminate hypotheses before accurately diagnosing the causes of symptoms.

upon system's operation. In case operational goals are not prioritized a-priori, the operator prioritizes them, identifying the ones that should be achieved first in the current situation. This prioritization determines the desired system state to be achieved (e.g., immediate restoration of malfunctions, immediate shutdown of the system, continuation of system operation, etc.). Then, the actions that need to be taken to return the system the desired state are identified. The next step is to determine the operational sequences and/or procedures for performing the identified actions. The last stage of the process is the coordination and continuous control over actions to be implemented.

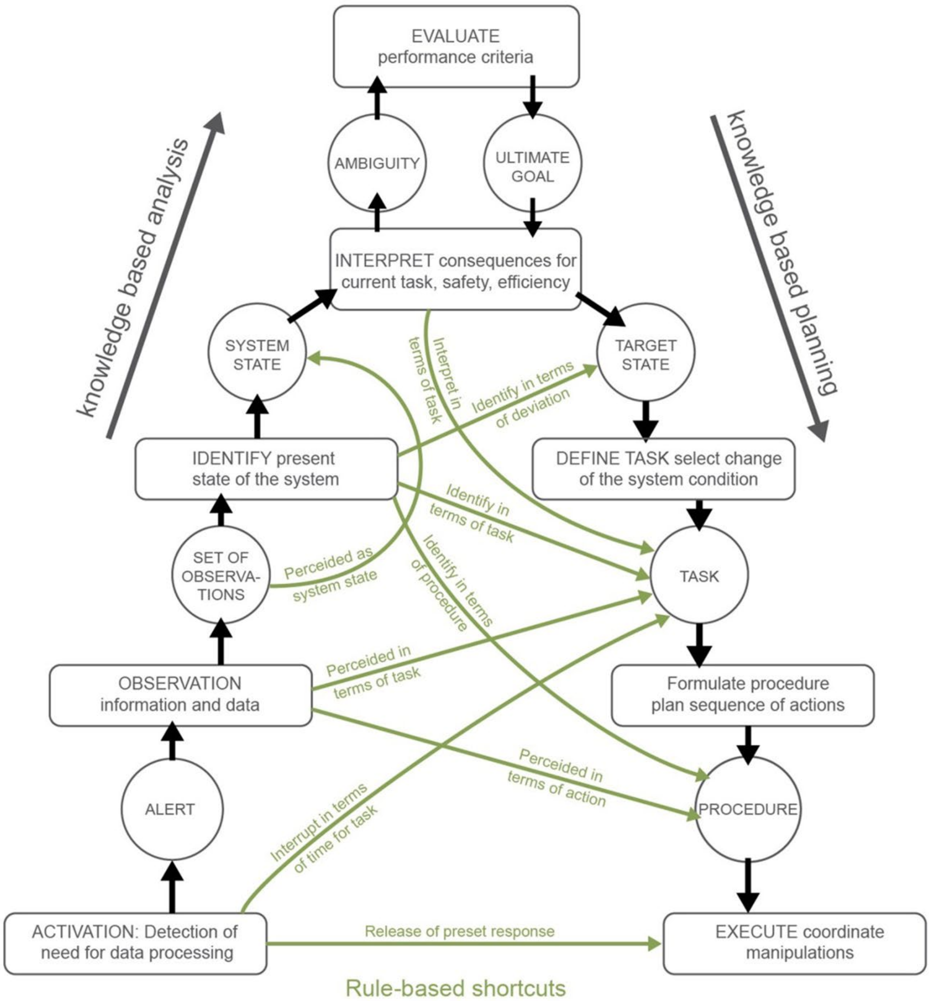


Figure 3.24. The "Decision Ladder" depicts both phases of cognitive activity for collecting and processing information (symbolized by a rectangle), and the knowledge stages of an operator (symbolized by circles) while troubleshooting a complex system.

The above description is of course only a normative description of the troubleshooting process. Indeed, a crucial property of the model is that it maps the common shortcuts or mental jumps in relation to the normative steps, depending on the skills of the operator or his/her familiarity with the malfunction. These shortcuts may range from skipping a single step to the total omission of the upper parts of the process, where as soon as the indications of the malfunction appear, the operator immediately takes actions to deal with it. The former behavior indicates an operator acting at a skill-based level of cognitive processing. If instead the operator, after collecting data on the malfunction matches it to a known troubleshooting routine, then this indicates that s/he has acted at a rule-based level of processing. In this way the Decision Ladder model maps the cognitive tasks undertaken by complex system operators into the Skills-Rules-Knowledge (SRK) taxonomy of cognitive processing.

3.6.5 Procedure following in complex systems

One might think that an essential aid for human operators called upon to deal with emergencies and malfunctions in complex systems would be to provide them with instructions –written or computer generated– on the procedures to be followed for each malfunction they may confront. In this way, their work would be greatly simplified, and their mental load and eventual errors reduced, increasing at the same time their efficiency. However, a series of studies of real work situations show that procedure following has its limits.

To begin with, in complex systems it is next to impossible to predict all possible malfunctions so as to formulate all corresponding procedures for their diagnosis and restoration. Research in complexity theory has shown that predicting the behavior of complex systems as a whole is impossible (Axelrod, 1997; Holland, 1992). But even if this could be done, given the extremely large number of possible malfunction situations, we again reach a problem-solving state in which the operator is called to select among an enormous catalogue of procedures in order to figure out which corresponds to the specific malfunction.

Also, in tasks where Standard Operating Procedures (SOPs) to be followed exist, it has been observed that operators often deviate from them, either because they feel that the specific situation could not be effectively dealt with these procedures, or because system information or necessary means are not available to them when they need them, or because they believe that it would take too long to implement them (we should not forget that operators are usually called upon to deal with malfunctions under time pressure and /or stress).

Another argument highlighting the limits of SOPs in dealing with system anomalies is that if a deterministic system could identify all its possible malfunctions and the respective diagnostic/restoring procedures, this, then, could be fully automated through appropriate software. But in how many real work situations has this been achieved? The presence of human operators in production systems that were originally designed to operate automatically, as well as the modern trend of using advanced automation systems to advise rather than replace skilled workers, are strong indications of the limits of SOPs.

Numerous field research studies over the last decades have demonstrated that complex system operators do not simply follow SOPs to get the job done (Carroll 1998, McCarthy et al. 1997, Orr 1996, Marmaras 1994). Of course, their activity is partly directed, partly constrained by procedures which translate objectives defined at higher levels of the work system. However, many issues remain to be resolved in-situ by line personnel through “rational” choices in the face of process or environmental variability. In fact, human operators face a multitude of contingencies and ambiguity in their day-to-day conduct, often having to invent workarounds, make judgments, decide and act under uncertain conditions. Work is, then, often accomplished in a dynamic distributed manner, in a collective, opportunistic and situated way (Hutchins 1995, Nathanael & Marmaras 1996).

The above arguments pinpointing the weaknesses of SOPs for diagnosis and troubleshooting of malfunctions, does not suggest that they are of no value or that any effort to improve them will be in vain. SOPs are invaluable as a common frame of reference and contribute decisively to coordinate the actions of those involved in the diagnosis and remediation of malfunctions. Furthermore, automatic control and decision support systems often contribute decisively to reducing mental load and improving the reliability of the whole system. The important thing is to recognize their limits and to acknowledge the contribution of skilled operators. Therefore, a balanced approach is recommended, with an effort to optimize the coupling between these artifacts and human skills.

3.6.6 A tale on procedure following... or not

Figure 3.25. A test launch of a US intercontinental ballistic missile, similar to the one assumed detected by the Serpukhov-15 satellite control center in Kaluga Oblast.

On 26 September 1983, in the early morning hours, the Soviet Union’s early-warning systems detected an incoming missile strike from the United States (Figure 3.25). Computer readouts suggested several missiles had been launched. The protocol for the Soviet military was to retaliate with a nuclear attack of its own. Thirty years later, duty officer Stanislav Petrov –whose job was to register apparent enemy missile launches– stated to the BBC:

> *"The siren howled, but I just sat there for a few seconds, staring at the big, back-lit, red screen with the word 'launch' on it," he says. The system was telling him that the level of reliability of that alert was "highest". There could be no doubt. America had launched a missile. "A minute later the siren went off again. The second missile was launched, then the third, and the fourth, and the fifth. Computers changed their alerts from 'launch' to 'missile strike" he says. "I had all the data to suggest there was an ongoing missile attack"*

Although the nature of the alert seemed to be abundantly clear, Mr. Petrov had some doubts. Alongside IT specialists, like him, Soviet Union had other experts, also watching America's missile forces. A group of satellite radar operators told him they had registered no missiles. But those people were only a support service. The protocol said, very clearly, that the decision had to be based on computer readouts. And that decision rested with him, the duty officer.

Mr. Petrov hesitantly decided not to report them to his superiors, and instead dismissed them as a false alarm. This was a serious breach of the procedures. His training was rigorous, his instructions very clear. The prudent thing to do would have been to pass the responsibility on, to refer up. But by overriding the procedures Mr. Petrov might have avoided a major catastrophe.

What made Mr. Petrov take this decision is not clear even to him. As he explained, it was a mixture of feeling and logic, just how strong and clear that alert was, its non-registration from auxiliary satellite radars, and his "gut reasoning" that one does not start a nuclear attack with only five missile strikes.

It is this unique ability of humans in combining such heterogeneous quasi logical, quasi circumstantial/contextual factors that makes them able to recognize rare critical events, not anticipated in formal system analyses, and take appropriate actions even by taking the risk of overriding strict procedures.

In the aftermath the investigation concluded that the false alarm was apparently set off when the satellite mistook the sun's reflection off the tops of clouds for a missile launch. The computer program that was supposed to filter out such information had to be rewritten.

3.6.7 Guidelines to support cognitive activity in complex systems

The question comes, then, on how to improve performance of human operators. Some hints have already been given in the section on cognitive biases. Here, the main guidelines are presented in a more thorough manner.

Perhaps the most essential aid is training of operators specifically targeted in providing the necessary knowledge for the formation of robust mental models. Such training should not only be limited in procedural matters but should also incorporate elements of the functional operation of the system, as well as its physical structure. Indeed, it is through integration of all three types of knowledge that robust mental models can be developed. Such models allow operators to effectively deal with complex unanticipated malfunctions, and devise original plans when no appropriate procedures exist. Finally, such training should also point to common cognitive biases that experienced operators could easily fall into. The above knowledge is best acquired through simulators, which offer the possibility for experiential training and familiarity in confronting rare malfunctions. Periodic refreshment and updating of such knowledge are also advised.

A second guideline is to design human-system interfaces explicitly aiming at supporting operators to detect and treat malfunctions, at the time and place they need it, in the most appropriate format. Effective interfaces should present system status at various levels of detail, from individual sensor readings to mimetic functional representations, up to configural displays that reflect the fundamental properties of the underlying work domain (e.g., energy or matter flows). Examples of human-system interfaces offering such configural representations of system parameters are presented in Chapter 8.

Finally, concerning the development of work instructions and SOPs:

- Their writing should be simple and at the appropriate level of detail. Indeed, very detailed instructions reduce effectiveness, since they fast become long and cumbersome, while very general ones do not offer substantial help.
- Their presentation must allow for quick identification of the appropriate actions to be taken, they must be easy to use and within the time limits available to operators. Towards this direction proper procedure indexing is key. An indexing per indications of malfunctions should be given priority over an indexing per system decomposition or other criteria (e.g., alphabetical).

As mentioned in the introduction of this section, designing to support human operators of complex systems requires an in-depth analysis of the work system, its environment, and requirements imposed to the operators at the cognitive level. This analysis, is followed by an analysis of how experienced workers cope with system anomalies (analysis of cognitive activities, skills and work practices), with the active participation of end-users themselves. Combining the results of the above analyses allows the designer to pose the following questions:

- Is there a more appropriate way of presenting information on system status for troubleshooting purposes?
- Is there a way to increase the reliability of the information presented?
- What supplementary information could be useful?
- Can information retrieval and processing be facilitated and how?
- Can we promote and facilitate the use of more effective strategies for diagnosing and correcting malfunctions, and how?
- Can mental load be reduced so as to mitigate poor cognitive performance (e.g., states of mental bias) or possible human errors and how?

Once these questions are tackled, the designer may, then, determine the most appropriate assistance means. Depending on the domain, such assistance can take many forms: designing appropriate human-machine interfaces, memory support systems, communication systems, specific supplementary training, etc.

References

Axelrod, R. (1997). *The Complexity of Cooperation. Agent-Based Models of Competition and Collaboration.* Princeton: The Princeton University Press.

Carroll, J. S. (1998). Organizational learning activities in high-hazard industries: the logics underlying self-analysis. *Journal of Management studies*, 35(6), 699-717.

Gibson J. J. (1977). The theory of affordances. In Shaw R., Bransford J. (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Erlbaum.

Holland, J.H. (1992). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, *Control and Artificial Intelligence. 2nd edition.* MIT press.

Hutchins, E. (1995). *Cognition in the wild*. The MIT Press.

Marmaras, N. (1994). Official and unofficial signals: a useful distinction for ergonomics. *Proceedings of the 12th Triennial Congress of the International Ergonomics Association,* Toronto: HFAC, Vol. 4, 328-330.

McCarthy, J., Healey P., Wright P. & Harrison M. (1997) Accountability of work activi-ty in high-consequence work systems: human error in context. Int. J. *Human–Computer Studies,* 47, pp. 735–766.

Nathanael, D. & Marmaras, N. (1996). Improving performance and safety in a complex work situation: Fishing manoeuvres aboard open sea trawlers. In A. Ozok & G. Salvendy (Eds.), *Advances in Applied Ergonomics.W.* Lafayette: USA Publishing Corp., pp. 1064-1067.

Norman, D. A. (1988). *The design of everyday things.* New York: Basic Books.

Norman, D. A. (2013). *The design of everyday things.* MIT Press.

Orr, J. E. (1996). *Talking about Machines: An Ethnography of a Modern Job.* Cornell University Press.

Rasmussen J. (1983a). Skills, rules, and knowledge-signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(3), 257–266.

Rasmussen, J. (1986). *Information Processing and Human-machine Interaction: An Approach to Cognitive Engineering.* Amsterdam: North Holland.

Rasmussen, J., Pejtersen, A.M. & Goodstein, L. (1994). *Cognitive Systems Engineering.* John Wiley & Sons, New York.

Reason, J. (1990). *Human error.* Cambridge university press.

Shingo, S. and Dillon, A.P. (1989) *A Study of the Toyota Production System: From an Industrial Engineering Viewpoint*. CRC Press, Boca Raton.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

Chapter 04: Needfinding

Chapter 04: Needfinding

Chapter Summary

In this chapter and onwards the focus is shifted from theory to methodology, introducing step by step the User Centered Design (UCD) process with relevant methods and examples. This first methodological chapter presents the most widespread methods of collecting information from prospective users that form the backbone of any UCD endeavor. The primary focus is placed on observation of human activity in naturalistic settings, with various means of externalizing cognition such as observable cues, eye-tracking, think-aloud or video assisted posterior verbalization. Other methods, such as laboratory based or sporadic sampling techniques (diary studies – photos – verbal reports) as well as questionnaires, interviews and focus groups are also presented. Original material from the authors' research and applied work are used throughout to substantiate the above. The User Center Design Methodology begins with the need-finding process.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of previous chapters.

4.1 User needs, unlike preferences, are often implicit

The evolution of automotive design has made cars safer, more efficient and more economical over the years. Progressive changes were made not only to the mechanical / electrical subsystems but also to the form of automobiles. For example, the hood ornaments that decorated many luxury cars nearly since the inception of automobiles (Figure 4.1), were gradually removed in the late 60's due to the increased injury risk to pedestrians in case of impact. Many other fittings like spinner wheel protrusions and external locks on the hoods were also banned for safety reasons while front bumpers got bigger and flatter.

Figure 4.1. A typical hood ornament. (Photo by Thom Quine under cc license).

The last protruding elements left to cars are the external side-view mirrors. Apart from being potentially dangerous to pedestrians or motorcyclists, they are quite fragile and functionally inefficient leaving blind spots and requiring the driver to take his eyes off the road to check them. Also, according to Hucho, W. H. (2005), traditional exterior side-view mirrors increase a vehicle's total aerodynamic drag by 2-7%, thus increasing fuel consumption and aerodynamic noise. Based on the above, the Alliance of Automobile Manufacturers along with Tesla petitioned the National Highway Traffic Safety Administration to allow cameras to be used in lieu of traditional mirrors (Tesla Motors, Inc. - Petition for Rulemaking, 2014). Many car manufacturers have proposed alternative design solutions in an attempt to suppress them. The most widespread idea is the replacement of these mirrors with cameras (Figure 4.2) that would cast the proper view on a display inside the car. That way not only do we get rid of these bothersome external fittings, but we also place this critical information inside the visual cone of the driver. Such camera solutions have already been implemented on luxury and sports cars and since costs are dropping, they are expected to also appear on conventional cars soon. Is this technological evolution though without its own problems compared to conventional side-view mirrors?

Figure 4.2. A camera replacing the side view mirror on a car.

In a field study concerning motorcycle driving in urban environment, we have investigated the perceptual strategies of experienced motorcyclists along with their points of concern using eye tracking technology (Nathanael et al., 2012). Data analysis revealed that one of the most prevalent elements of concern was the side mirrors of cars being overtaken (Figure

4.3). Interestingly, their concern was not mainly to avoid hitting these protruding elements but to check if they have been spotted by the driver or to infer her/his imminent intentions. In a retrospective think-aloud analysis of the eye tracking videos, the motorcyclists confirmed that this information is rather crucial during overtaking and that they would feel quite insecure if it was missing.

Figure 4.3. A frame of the eye tracking study, showing the focusing of the motorcyclist on a side view mirror to understand the intentions of the driver.

Under this light, side-view mirrors are not only essential for the car driver, but also a source of safety-critical information for any adjacent motorcyclist. Therefore, if external mirrors were to be eliminated from future cars, alternative means should be developed to provide motorcyclists with the relevant information. The lesson here is that when designing for the future, we must be vigilant not only to formal system uses but also to informal, tacit or even deviant ones. We must recognize all user groups involved, and conduct a detailed analysis of their needs, habits and skills so as to predict how these would be affected by a new design. The process employed for such analyses is called needfinding. The need-finding process, along with its dedicated methods and techniques, constitute the backbone of User-Centered Design.

The theory behind needfinding was developed by Robert McKim at Stanford University in the 1970s as part of the theoretical foundations of design thinking. Unlike other user research methods (e.g., in marketing), needfinding explicitly downplays user "wants" and focuses mainly on their needs. Needs are difficult to articulate as they are often implicit. Therefore, needfinding is explorative as a process. We look without first having a clear picture of what we are looking for, and trust that user needs will emerge during the process itself. The main purpose is to discover latent user necessities, gain empathy for them, understand their motivations, and identify expert user's techniques/tricks that could point to innovative directions for future system design.

4.2 Analyzing human activity: a complementarity of views

To explore the needs of people engaged in a specific activity, we have to understand what they do and why. An obvious way to start is by asking. If we indeed ask someone to describe what they do, they may readily provide a description. Such descriptions typically contain the motivations, main goals, and normative task sequences, (i.e., what can easily be spoken out by the acting persons), which are the so-called declarative parts of the activity (see Chapter 2). However, these descriptions will be subject to common biases; they typically omit certain aspects as too complex, difficult to articulate or inappropriate to disclose, while overemphasize others as more socially suitable. Moreover, as exemplified in Chapter 2, people cannot verbally articulate all that they know or do. Therefore, even in thorough description efforts, it is inevitable that embodied parts of the activity (i.e., skills) will be missed. The above observations drive us to an important point for the need-finding process: what people think or say about what they do, does not equate to what they actually do in practice.

When the activity being studied is embedded in a larger system or domain, then the analysis becomes more complex. In these cases, it becomes essential to be fully aware of all the regulatory and organizational requirements that influence the activity under study. This is most evident in professional work-systems. A work-system is characterized by a goal structure reflecting system designers and/or higher management vision, as well as prescriptions at many levels of abstraction. System designer/higher management vision and intent is typically an idealistic view of what people are expected to do and held accountable for, without considering the messy details of reality. Prescriptions include all the explicit rules, e.g., laws, regulations, procedures, standards, job descriptions, with which the work must be aligned. They, thus, strongly influence work activity by exerting various constraints and pressures towards compliance.

Consequently, in activities embedded in larger systems it is essential to analyze not only what is described and practiced by the actors themselves but also what is imagined by higher management aspiration and intent and prescribed through the system's regulatory framework.

Following the above observations, Shorrock (2016) has distinguished four views of human work activity; these are briefly described below:

Work as Imagined	The work as imagined by people outside of the work, e.g., managers, engineers, or even the public. This is a distorted and over-simplified view of work typically assuming optimal resources, predictable outcomes, no deviations from rules, always prioritizing according to the system goals.
Work as Prescribed	The formalization or implementation of work-as-imagined through rules, procedures, checklists, job descriptions, etc. This written modus operandi is often under or over specified and essentially impossible to follow at all times in practice.
Work as Espoused (or disclosed)	How workers think about or describe their work. This view is a "rationalized' account of their experience, or a version of work as prescribed or imagined, depending on what is expected by the person(s) asking and of possible consequences of disclosing it.
Work as Done	The actual work as experienced during action, full of tradeoffs and compromises, variability to get the job done, conflicting goals, suboptimal resources, disregard of procedures and people finding solutions on the fly.

The relation between the views is illustrated in Figure 4.4, mapped according to perspective (i.e., system or actor) and substance (imaginary or implemented). As the figure suggests, the four views of human work do usually overlap, but not completely, leaving areas of commonality, and areas of difference.

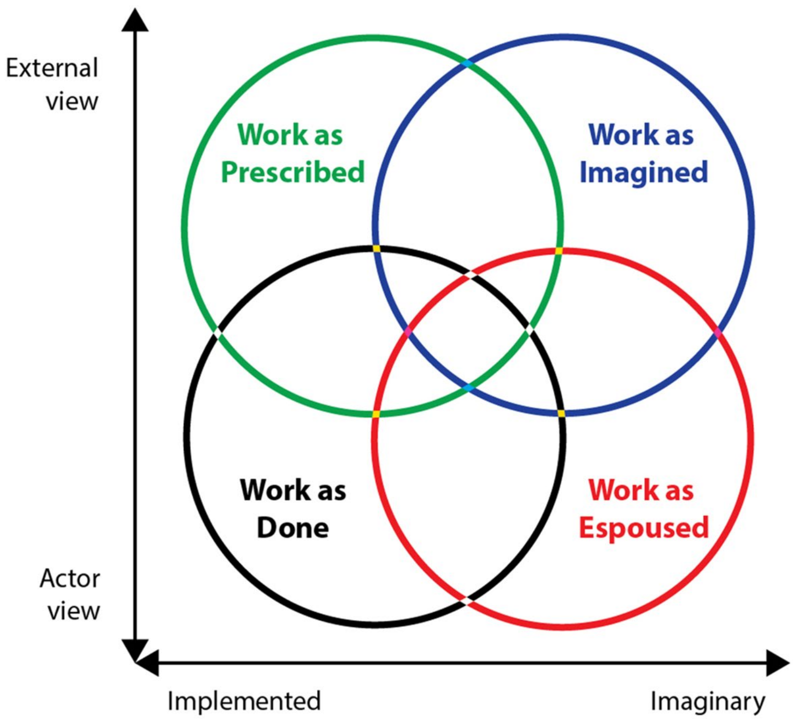


Figure 4.4. The different views of work mapped according to perspective (external or actor) and substance (imaginary or implemented).

Exploring the differences between views is a good starting point to reveal hidden user needs. Disparities between “work as espoused” versus “work as done” hints to unacknowledged or unarticulated needs of the workers themselves. For example, whenever a worker or user of a system is unable to provide a consistent cause-effect relation between their actions and outcomes, serves as an indication of a poor mental model in need of improving through training and better interface design. In a similar manner, specific misalignments between “work as prescribed” versus “work as done”, e.g., when workers constantly deviate from a formal procedure, indicates an unarticulated need for changes in work methods, resources or tools.

Finally, in new system design, work as imagined serves well in reminding a designer or an engineer that their expectations on the future user activity can only partially encompass what will happen in practice. This limited ability to foresee user activity with a new system creates the need for iterative testing of prototypes at various stages of system development, as we will see in the next chapters.

When conducting a work analysis, our objective is to understand work activity in all its complexity. This includes what is done, how it is done, but also why it is done, hence the motivations, goals, constraints, strategies, behaviors and beliefs of the people involved. All the above elements are necessary in order to reveal real needs. Therefore, the analysis of work cannot be limited to work as prescribed or simply to a description of work as done. Similarly, it cannot be limited to work as we imagine it, or work as people talk about it. Only by considering all four of these views, can we hope to get a firm grasp of what’s going on and why.

4.3 Direct ethnographic observation

In June 1914, Bronisław Malinowski (1884-1942), a Polish-British anthropologist departed London, travelling to Australia, to conduct an ethnographic study on the tribes that resided the islands of Papua. His journey to Australia was supposed to last only about half a year, as he was mainly planning on attending a conference there. Shortly afterward, his situation became complicated due to the outbreak of World War I; although Polish by ethnicity, he was a subject of Austria-Hungary, which was at the state of war with the United Kingdom. Malinowski, even though he was at risk of internment, he decided not to return to Europe from the British-controlled region and after intervention by a number of his academic colleagues, British authorities allowed him to stay in the Australian region. What is more, they provided him with new funding.

His first field trip, from August 1914 to March 1915, took him to the Toulon Island (Mailu Island) and the Woodlark Island. This field trip was described in his 1915 monograph “The natives of Mailu”. Subsequently, he conducted research in the Trobriand Islands in the Melanesia region. He organized two larger expeditions during that time: from May 1915 to May 1916, and October 1917 to October 1918, in addition to several shorter excursions (Figure 4.5). It was during this period that he advanced the practice of participant observation, which remains the hallmark of ethnographic research today.

Figure 4.5. Bronisław Malinowski surrounded by natives of Papua New Guinea.

What differentiated Malinowski from his colleagues at that time was his extensive stay among the people he studied and his gradual integration into their community. Shortly after, he started noticing and understanding practices, behaviors and habits that were invisible to the untrained eye, or simply would not take place in front of a stranger or an external observer. Apart from that, he started understanding their values, their incentives, and more importantly he started seeing the world through their eyes. Malinowski emphasized the importance of detailed participant observation, and argued that anthropologists must have daily contact with their informants if they are to adequately record the "imponderabilia of everyday life", which are considered of the utmost importance to understanding a different culture. He stated that the goal of the anthropologist, or ethnographer, is "to grasp the native's point of view, his relation to life, to realize his vision of his world".

Malinowski's approach was later adopted in Human Factors & Ergonomics studies by various researchers who emphasized the role of field observation for identifying user needs in technological systems [see, for example, the work of Lucy Suchman (1987) in computing systems, or the work of Edwin Hutchins (1995) and Hutchins & Klausen (1996) in maritime and aviation domains]. Such "micro-ethnographies" of worker activity have since become the primary method of inquiry in most HF/E studies (see Daniellou & Rabardel 2007; Garigou et al., 1995). By mingling with, and systematically observing prospective users or workers performing tasks in their natural environment, we are able to collect rich data on their skills and needs. When conducted properly, field observation provides the richest picture of how work is actually done, away from unjustified interpretations, biases and mediating channels. Also, the experienced observer can elicit elements of work that are either tacit or implied but never explicitly expressed.

Technically, when conducting filed observations, we are searching to gain insight on:

- what people's goals, values, motivations and beliefs are,
- what their competences (i.e., their skills and know-how) are,
- how context (e.g., environmental or system variability) affects their work,
- how people differ among them in the above,
- how their activity is integrated in a larger setting.

Once we get a fair understanding of the domain in question, we continue with more in-depth analysis of people's activity through detailed/ structured observations. Specifically, through "activity analysis" we can identify:

- typical workflows that are usually followed to perform the job in both normal and emergency situations,
- possible ways that the operators have "invented" to facilitate the execution of their work and to deal with the difficulties they encounter (e.g., informal signs that they use),
- aspects of the cognitive skills and mental models that they have developed,
- dilemmas and adversities encountered while performing their work,
- possible errors that occur.

Conducting activity analysis is a form of art, requiring the combination of technical, theoretical, as well as social skills, which can only be mastered through practice. Suffice it, here, to briefly present some essential guidelines:

- build confidence and cooperative climate with the people you are observing,
- take the role of the apprentice not that of the inquirer,
- learn all/most workflow sequences and understand why,
- aim to find tricks of the trade (e.g., informal practices),
- validate practices by discussing them with people (not the other way around!),
- be attentive to self-made work artifacts (post-it notes, special tools, ready-made solutions) (Figure 4.6.),

- try to get access to possible failures (through log files, reports, etc.).

Figure 4.6. Post-it notes and annotations on a remote-control station of the Hellenic Railways Organization (O.S.E.), Thessaloniki, Greece. Such operator interventions provide direct insights into how a system is actually used and the information that the system is lacking. (Photo by Konstantinos Tsakalidis)

Systematic observations of activity should be performed in a way that reduces the "observer effect" as much as possible, i.e., the inevitable influence on people's behavior from the fact they are being observed. Observance to the above requirement ensures the "ecological validity" of an observational study (ecology here is used in the sense of the totality of determinants affecting participant behavior, including motivational, ethical, behavioral and environmental factors). However, depending on the specifics of the work being analyzed, certain reductions can be made in the above requirements. For instance, when we want to gain insight in worker behavior in exceptional situations or in safety critical scenarios (e.g., loss of primary coolant in a chemical plant, engine failure in an aircraft), we may do so through simulated parts of the work system, both in-situ or in the laboratory. Equipment typically used for recording worker activity include audio and/or video recorders, eye tracking devices, screen capture software, etc. However, activity recording can also be done simply with pen and paper if the observing task is reasonably paced with few points of interest over a period of time (see Figure 4.7 for a typical pen and paper observation grid and transcript of operator's activity timeline). On the other hand, video recordings are almost indispensable when the task is highly paced, rendering actions or decisions impossible to capture in real-time (e.g., athletes playing table tennis). In such cases, the observation is mainly conducted posteriorly to the actual task, several times along with the presence and aid of the participant/s themselves.

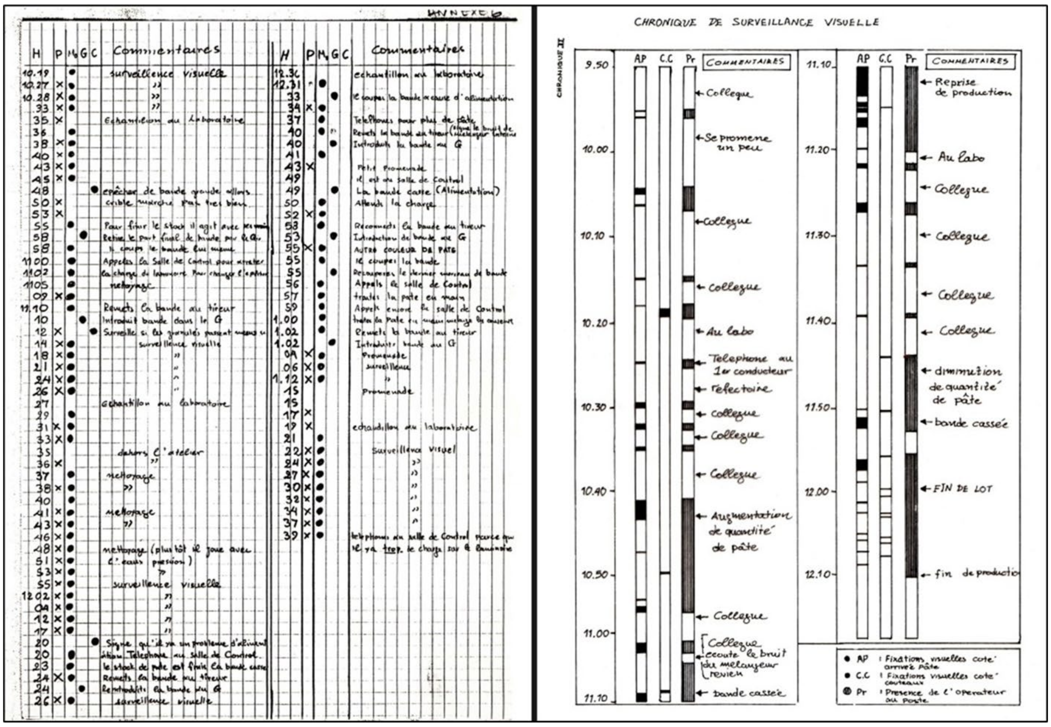


Figure 4.7. Example of activity analysis extract; structured observation grid (left) and transcript (right) of operator's activity timeline in a chemical plant.

4.3.1 Where knowledge resides and how to capture it

Jack Whalen, a social scientist, while he was working at Xerox's Palo Alto Research Center, he had spent some time in a customer service call center studying how technicians used a specific software that was supposed to help employees tell customers how to fix copier problems (paper jams, etc.), by matching descriptions of a problem against a knowledge base of known solutions (Whalen et al., 2002). The trouble was that the employees were not using it. Management, therefore, decided workers needed an incentive to change. To this end, they held a contest in which workers could win points (convertible into cash) each time they solved a customer problem, by whatever means. The winner was a veteran named Carlos, who scored more than 900 points. Carlos really knew his stuff and everyone else knew this, too. But Carlos never used the software. However, the runner-up, named Trish, was a shock to everyone. She had been with the company only a few months, had no previous experience with copiers, and did not even have the software on her machine. Yet her 600 points doubled the score of the third-place winner. Her secret: she sat right across from Carlos. She overheard him as he talked, and she persuaded him to show her the inner workings of copiers during lunch breaks. The solution to the customer service support was not the incentives to use the existing software but how to turn Carlos' know-how into a knowledge base.

The story above illustrates two important points to keep in mind when conducting observations:

- How, at the start of observations, one should restrain from fixed presuppositions on where problems lie, but to genuinely observe and analyze work as it really happens with all its contextual peculiarities.
- How crucial it is to identify and collaborate with subject-matter experts in the observation process to get fast/true insight into the work and understand its various aspects.

We know that experts have tons of knowledge on the systems they work with, but most of it is tacit or not easily expressed out of context. In this sense, the analyst's role is to recognize where knowledge resides and try to extract it through proper engagement and collaboration with the right people. Identifying subject-matter experts is often quite easy; genuinely collaborating with them is not. People will not invest time and share what they know freely, just because they were instructed or asked to do so. To truly make the effort to share their know-how, they must understand the benefits; it must be a deliberate choice. A good way to achieve this engagement is to give them a formal role in the research/design team, guarantying that they will have influence on the obtained results.

4.3.2 Verbalizing one's cognitive activity: the Think-Aloud method

When trying to capture the cognitive aspects of an activity, the mere observation or recording of physical actions can only give us clues to make assumptions (e.g., direction of gaze, pressing of a button). One common way to access what happens in the mind of a person while working, is to ask them to spontaneously verbalize their thoughts in real-time. By listening to a person's externalized thoughts while performing an activity, we can access their momentary goals, intentions and interpretations as they happen, without the simplifications and post-rationalizations that are typical of responses given through posterior interviews.

The "think-aloud" method is well suited for tasks that entail cycles of decision making, planning and interpretation of the received feedback. The main drawback of the method is that concurrent verbalization may interfere with the performance of the actual task; for this reason, it is generally not recommended for intensive, time-constrained and/or highly paced tasks.

Probing people to externalize their thinking while at work is not always straight forward; some people may find it quite natural while others may not. However, various tricks can be employed to this regard. Here is a well-known example from the literature: In 1983, while working at XEPOX PARC, Lucy Suchman conducted a micro-ethnographic study to evaluate a novel copy machine interface (Suchman 1987). To do so, she came up with an original idea. She asked two of her colleagues to try setting-up the machine together to perform a complicated copy task, while video recording them. The video and audio transcripts proved exceptionally rich in capturing the problem-solving processes followed by the two participants. The decision to use two persons instead of one proved ingenious, because, in this way, the verbalization of their thoughts was done spontaneously while trying to cooperate. Nowadays, this study is considered as one of the hallmark instances of corporate ethnography.

4.3.3 Recording people's behavior: The eye-tracking technique

Various technical means can be used or especially crafted for monitoring people's behavior, depending on the project needs, from camera/voice recorders to geolocation devices, to wearable sensors for movement, muscle activity, or even electroencephalography (EEG) for measuring brain activity. It should be noted, though, that such technologies tend to obstruct or otherwise interfere with the user's naturally occurring activity, and thus, should not be used gratuitously. In this section, the use of eye-tracking for needfinding will be presented, as it is a technology that can be employed in various environments and offers a particularly privileged view of users' activity, i.e., of seeing through their eyes.

The recording of a person's point of gaze (i.e., where one is looking) can be achieved with a family of devices known as eye-trackers. The main advantage of the technique is that it produces a detailed recording of all the points of visual attention of a person, be them deliberate or not[8].

This technique can produce a variety of visualizations such as:

- still scene scan-paths (or gaze-plots),
- video scan-paths (or Gaze replays),
- heatmaps.

A wide range of disciplines use eye-tracking techniques, including Cognitive Science, Psychology, Human-Computer Interaction, Marketing

[8] The most widely used current designs are video-based eye-trackers. A camera focuses on one or both eyes and records eye movement with a sampling rate of 50-100HZ, as the viewer looks at some stimulus, either a visual screen or the physical environment. Most modern eye-trackers use the center of the pupil and infrared/near-infrared non-collimated light to create corneal reflections. The vector between the pupil center and the corneal reflections can be used to compute the point of regard on a flat surface or the gaze direction. A simple calibration procedure of the individual is usually needed before using the eye tracker. Along with the pupil, the stimulus is also recorded, either through another camera or by screen capture software where appropriate. Then, the eye movement data are superimposed on the stimulus recording. In this way, we can have a representation of the field of view of the participant along with eye-fixations and the saccade pathways (i.e., eye's movement) between them.

research and Medical research. In Human Factors and Ergonomics, eye-tracking is used for both need-finding observations and prototype evaluation, depending on the objectives and phase of each study.

An example of eye-tracking use

In a field study conducted in our laboratory, we used a head mounted eye-tracker to record naturally occurring driver interactions with other road users at high density urban intersections (Figure 4.8). We, then, showed to each participant driver his/her eye-gaze video recordings asking them to comment on their behavior; a technique known as video-assisted retrospective commentary (Nathanael et al., 2018). The main goal of the study was to gain insight on how drivers coordinate with other road users to solve traffic conflicts and ambiguities on the road. Specifically, the study aimed at identifying cues and signs used among drivers and pedestrians to coordinate their path planning, in an effort to devise features that autonomous (i.e., driverless) vehicles of the near future should implement for communicating with human drivers.

Figure 4.8. A driver wearing eye-tracking glasses.

Participant drivers were instructed to drive their own vehicle in their normal style at a selected course. The driving duration was estimated to approximately 15 minutes. Immediately following the driving session, participants returned to the lab and were asked to watch their eye-gaze video recording, while commenting aloud on their behavior for each case of interaction with a pedestrian or other vehicle (Figure 4.9). This retrospective analysis was necessary as perceived cues, interpretations and decisions taken are not easy to verbalize while driving due to the complexity, criticality and fast pacing of the driving task. Moreover, thinking aloud while driving would interfere with, and alter the participants' natural driving behavior. While watching their own visual behavior through the video, pausing and re-playing it as they liked, the drivers were asked to comment aloud on their lived experience and reflect upon their decisions.

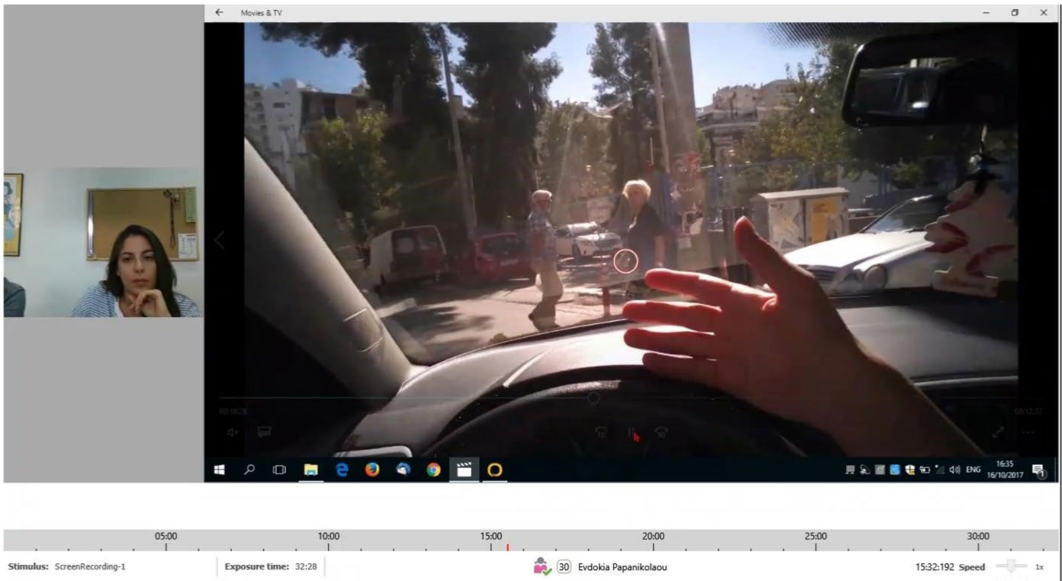


Figure 4.9. A participant of the eye-tracking study, commenting on her decision making while watching her eye tracking video recording.

As noted above, this is a need-finding technique, where the analyst does not have an a priori hypothesis of what s/he is looking for but, instead, s/he tries to elicit needs by studying the users' activity (e.g., goals, practices, behaviors, thoughts). By this approach, a designer can, then, think of alternative ways of interaction with a robotic driver, where the same needs must be satisfied through a different medium.

On the methodological side, it is important to note that watching one's own eye-gaze after performing a visual task, significantly increases our ability to recall details of the lived experience and comment on it, especially if the retrospective commentary is done immediately after the performance of the task.

Eye-tracking can also be used on detecting the eye-gaze on a monitor screen using a slightly different technological apparatus, but based on the same principles. While a participant performs a task in front of a monitor (e.g., an online shopping task), the eye tracking device at the base of the monitor remotely reads the pupil of the participant's eye and determines the direction and concentration of their gaze. Dedicated software, then, generates data about these actions in the form of heat maps and saccade scan-paths.

Eye-gaze visualization outputs

Various visualization outputs can be produced from eye-tracking data through post-processing via specialized software, the most common being scan-paths and heat-maps.

Scan-paths trace the eye's movement between areas of focus (Figure 4.10). Eye movement is characterized by series of brief periods of focused attention to specific points (called fixations) and rapid eye movements between them (called saccades). After software processing, a scan-path shows a series of circles indicating fixations and lines connecting them that indicate saccadic movements. Typically, cycle diameter indicates the relative duration of each fixation.

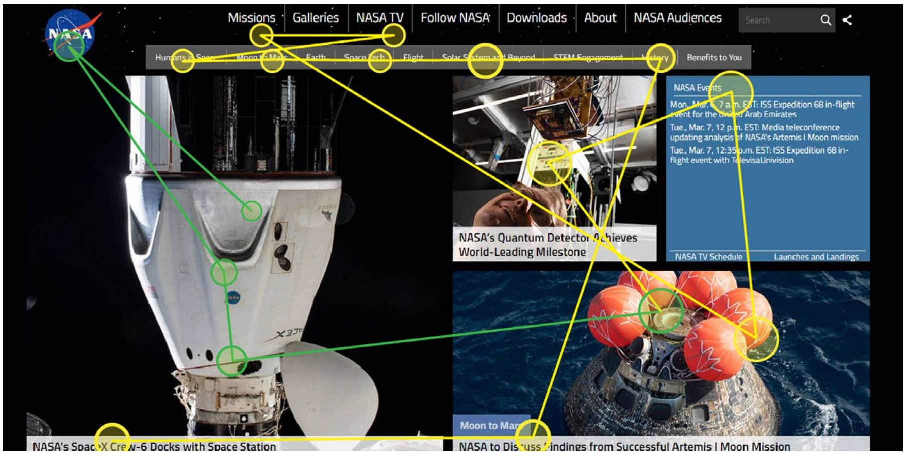


Figure 4.10. A scan-path on a web page. The green part consists of the first five fixations while all the rest are yellow.

Heat maps represent where participants concentrated their gaze and for how long they have gazed at a given point (Figure 4.11). Generally, a color scale from cyan (low) to red (high) indicates the concentration of focus. Thus, a red spot over an area of a webpage indicates that a participant, or group of participants, focused on this part of a page for a longer period of time. Heat maps are generally summative pictures of gaze behavior per group of participants rather than per single ones.

Figure 4.11. A heat-map of the Athens Metro homepage (summative recording of 5 participants).

4.3.4 Log files and existing records

All human activities leave traces, and when these activities are part of formal procedures, it is more than likely that some traces are recorded and archived inside the organization. Such records may be:

- Formal conversations between operators at different hierarchical levels. These conversations can either be recorded as part of meetings or through some team collaboration platform (e.g., air traffic control – pilot conversations or company agent – customer conversations).
- Text messages and reports that are also addressed to colleagues, superiors or external collaborators.
- Help-desks questions and answers of users of specific systems or devices.
- Log files of information technology systems, (i.e., big data analytics), from which one can collect rich data on user actions, such as frequency and quantity of various requests, exchange of messages with specific recipients, etc.
- Analysis of recorded errors during task execution, i.e., actions that do not lead to the desired outcome.

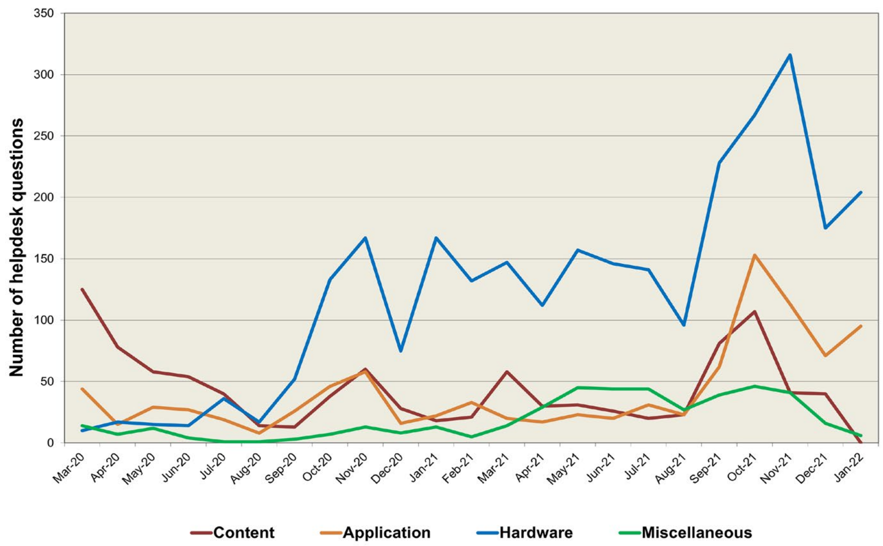


Figure 4.12. Analysis of helpdesk questions indicating frequency and evolution per topic. Such analyses help focus on the most prevalent issues before conducting more thorough content analysis.

Various professional or amateur communities maintain specialized online forums in social media platforms (e.g., couriers, photographers, users of specific software). Such forums contain a huge amount of user generated content and can be an invaluable source of information about user needs, even prior to going to the field or meeting prospective users. Reading through endless pages of such content can be a copious job but the advantages it offers easily outweigh the effort. The avid analyst can get unmediated insight into social elements such as group ethnography, concerns and attitudes to specific pain points that directly inform design ideation. In addition, a quick content analysis of the frequency of specific topics or even specific terms may even provide quantitative indices of their relative importance for the concerned community. Finally, an analyst can even use such forums to get feedback from users on specific questions or to use such platforms for recruiting participants in formal need-finding surveys (see later in the present chapter).

To overcome the analysis of a potential huge amount of data, Artificial Intelligence (AI) content analysis and composition tools can be used to assist in interpreting user-generated content. Such tools can be used to identify patterns in user language and behavior and even create natural and engaging prompts for users, leading to more insightful and detailed feedback. Such AI tools already have many features that a designer can rely on, such as ability to generate coherent and contextually relevant responses to user queries. However, there are some limitations that must be considered. For instance, such AI tools are not yet capable of understanding nuances or context that a human researcher might pick up on. Also, their trustworthiness depends on how they are used and trained. It is, thus, crucial for the researcher to have a general understanding of how such tools function and where their sources of information come from to avoid biases and ensure that the original data are trustworthy. It is, also, imperative for the researcher to supplement AI derived data with additional research methods in order to ensure validity.

Finally, it should be noted that in professional domains, the record keeping used for worker monitoring may be deliberately manipulated to reflect work as prescribed. In a similar manner, anonymous online user forums tend to reflect the work as espoused and not necessarily how it is actually done. However, the anonymity of such media tends to allow people to be more open to admit deviances and irregularities than in formal interviews.

4.4 Survey methods

In 2009, the American retail corporation Walmart, in an attempt to outrival its basic competitor who had a cleaner look in their stores, conducted a survey on its customers' opinion, asking them to answer the following question:

"*Would you like Walmart to be less cluttered*?"

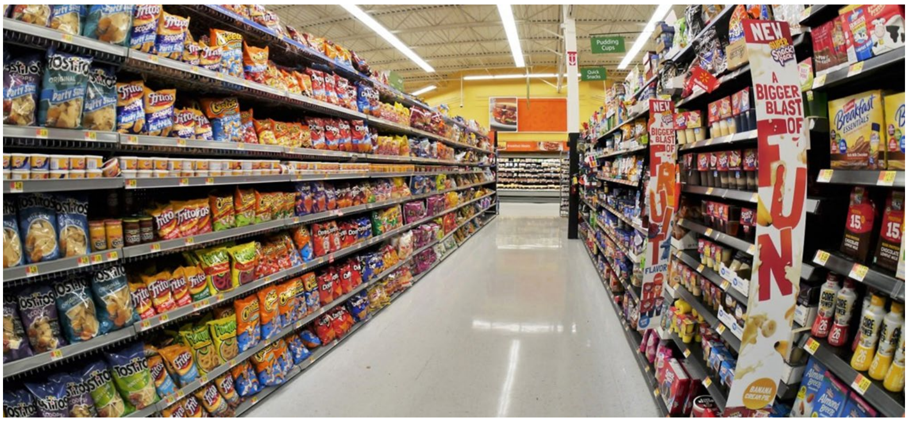


Figure 4.13. Arrays cluttered with potato chips and other snacks in Walmart, Wenatchee Washington. (Thayne Tuason Wikimedia Commons)

The respondents' answer was an overwhelming "*yes*", directing the company to spend millions of dollars to clear out space, removing 15% of inventory and shortening shelves on their stores (Figure 4.12). After implementing the changes, to their surprise, same-store sales plummeted, by about $1.85 billion, and Walmart decided spending some more millions to undo what they had spent hundreds of millions doing. But why did this happen? Wasn't all this done based on customers' request after all? Truth is Walmart wanted to seem like it was listening to their customers, but they were not actually attending to their needs. "Walmart came up with the answer first, then asked customers to agree to it," writes the Good Experience blog. This is the peril of listening to what you want to hear instead of attending to what your customers' actual practices show.

As we will see next, Walmart's question was flawed for two reasons: It was both hypothetical and leading. It is rather common to fall into such traps where instead of exploring actual people's needs, we are trying to validate our own ideas of what is best for them through surveys.

Survey methods in user research include all alternative ways of collecting information by asking users to express themselves, e.g., questionnaires, interviews and the like. Surveys are generally considered less reliable than direct observation methods for answering questions on people's behavior, but they are often necessary for collecting data on people's feelings, attitudes and preferences. When used in professional settings so as to gain access to work activity, it must be taken into account that the information we get is probably the work as espoused by them or as prescribed by management, and not as it is actually done. To overcome this challenge, it is critical to complement this method with field observations and log files analysis, and remain as close as possible to the basic principles that are presented below. Nowadays, the popularity of the survey methods lies in three parameters; they seem easy to design, they provide the ability to collect data from a large sample of people, and the data collected can be easily statistically processed (which, however, often provides more of a "scientific scent" rather than substantial insight).

4.4.1 Basic principles of survey design

Although designing a survey (in the form of interview or questionnaire) seems easy at first, in practice, good survey design requires knowledge, effort and considerable preparation. This section will present some basic principles and guidelines, but it cannot replace the knowledge provided by special manuals (see for example Oppenheim, 2000) to which the reader is referred for further study. Apart from the above, the reader should have in mind that most of these textbooks have been developed for areas such as Marketing, Psychology, Sociology, etc.; therefore, our treatment of the subject will particularly concern survey design related to user-needs finding for design purposes.

Before deciding on a particular survey method, we must clearly define what kind of information we need and why. Explicitly stating what we want to learn and why, preferably in a tabular format, helps us disambiguate between the two (i.e., data to be collected vs what we will infer from it). Once we accomplish this first step, we, then, need to find out how we will collect this information. At this stage, we need to check if the information we need is already available elsewhere. For example, there is a large amount of public data available from statistical offices, public opinion research companies, special journals, etc., that is much more valid and reliable than what we can collect from a limited study on our own.

To decide on the collection technique (i.e., interview or questionnaire) and on the specific wording of questions, we need to ask ourselves the following:

- Do the people we address have the information we will ask for?
- Can they answer our questions?
- Are they willing to answer honestly?
- Can they express in words what will be asked of them?

For instance, suppose we need to assess how dangerous a workplace is. By directly asking the employees: “How dangerous is your workplace?” will the answers be reliable? Since each person has different experiences and understands the concept of safety in a unique way, it is highly unlikely that we will get useful answers. To get a more reliable answer to our question we should rather ask the employees something concrete, e.g., if they have witnessed or been involved in an accident, or what type of dangers they recognize in their workplace. Therefore, to get the information we need (the “what”), we have to transform it into something that people can answer based on their actual experience (the “how”). It is up to the researcher to infer the information s/he needs based on these responses. Table 4.1 provides examples of (i) information sought, (ii) justification of why we want it, and (iii) transformation of what we want to learn into answerable questions.

Table 4.1. Examples of information sought, justification for it and how to ask for it.

What	Why	How
How often does a certain user group consult the weather	To decide where to place a weather bulletin on an interface	Ask the question: how many times have you checked the weather in the last three days?
What types of programs one uses on his/her washing machine	To decide on the options of a new washing machine interface	Ask the question: check the boxes of temperatures / programs you currently use
How experienced a motorcycle rider is	To decide if one is expert on the task or not	Ask the questions: How many miles do you travel per year? Since when do you hold your license?
How does one use his home heating system	To decide how to communicate a more ecological approach when programming the heating	Ask the question: In winter do you usually keep a constant temperature, or you alter it during the day?

In case of doubt about the ability of respondents to provide reliable answers to some questions, it is advisable to look for other methods in order to gather the necessary information. For example, it is unadvisable to ask hypothetical questions in user needs' surveys. Hypothetical questions are typically used to elicit opinions and beliefs about imagined situations or conditions that do not exist. As such questions are based on supposition and not facts, their validity for design purposes is very low.

At this point, it is worth mentioning a famous saying by Henry Ford (1863-1947), who, on the occasion of the design and production of the Ford Model T that appeared on the American market in 1908 as the first mass produced car aimed at middle- and low-income people, said: "If I had asked people what they wanted, they would have said faster horses".

4.4.2 Questionnaires and Interviews

Usually, interviews are preferred when at an exploratory stage when we need initial in-depth information from the future users of our system. In order to conduct one, before meeting the participants, we just need to have decided on the topics we want to collect information on (called interview points). We are thus able to ask questions on the same topic in many ways, until we are sure that the respondents have understood what we want to know. Also, from their expressions and reactions we can understand if they are able to answer, how sure they are about the accuracy of their answer, etc., in order to reformulate the questions or give them helpful examples. Depending on the level of questions preparation and protocol following during their conduct, interviews can either be structured, semi-structured or free. Structured interviews are preferred when specific issues need to be addressed, while free ones are best used when first exploring the design scope. Semi-structured interviews lie between the two, leaving more space in protocol following than structured ones to address emerging issues. The analysis of the interviews' outcomes can be conducted through Thematic Analysis which will be discussed in Chapter 5.

On the other hand, questionnaires are usually conducted to statistically validate certain assumptions or to make the respondents choose between alternative options. The questions can either be open-ended, where the respondents can formulate their answers as they wish, or closed-ended, where they must choose between predefined answers. While open-ended questions provide much richer answers, without limiting the respondent to predefined options, they are more difficult to be processed as they require content analysis. The choice between the two types depends on what information and data we need. A common practice is to form a pilot questionnaire with open-ended questions and distribute it to a small audience. Depending on a summative analysis of the answers that we will receive, we can first design the final closed-ended questions and then, distribute them to the actual audience.

Good practices when designing questionnaires

- A pilot phase where an initial questionnaire is administered to a limited number of responders is almost imperative so as to weigh the rating or to check on the validity of alternative answers that you will provide to closed questions. It also helps rephrase questions that the respondents found difficult to understand. For this reason, pilot questionnaires are better completed in the presence of the researcher.
- There should always be a brief introductory note explaining the aims of the study, the estimated completion time and the details of the researchers.
- The questionnaire should be attractive to the respondents, making sure that the questions that will stimulate their interest are placed at the beginning.
- Questions concerning demographic data (e.g., age category, level of education) should always be justified for the proper study; asking such questions (e.g., gender) without justification not only makes the questionnaire longer, but also hinders the researcher's ability to focus on pertinent issues. Unless used to split the user group into alternative questionnaire paths, demographic questions should be placed at the end of the questionnaire, as they are the least attractive for a respondent.
- The questions should be as simple as possible and "speak the language" of the respondents, using as much as possible the terms and vocabulary they use in their daily lives.

- The wording of the questions should be polite trying not to offend the respondents in any way. For example, "*do you always remember to pay your energy bills on time?*" is preferable to "d*o you pay your energy bills on time?*"
- Questionnaires should be brief (preferably taking less than 10min to complete) and should only seek the information absolutely necessary for the study.
- The questions should be about the actual lived experiences of the respondents and their goals, avoiding opinions based on assumptions.
- Questions that require recall from recent memory are preferable to general frequency questions. For example, "*How many times have you been to the movies in the last month?*" is preferable to "*how often do you go to the movies?*" or "*how many coffees did you drink yesterday?*" is preferable "*how many coffees do you drink a day?*"
- In closed-ended questions with YES/NO answers, negative wording of the question should be avoided (e.g., "*when I encounter difficulties in my work, I do not consult my colleagues*").
- The researcher should always keep in mind that the answers should be made easy to index and process. For example, keeping the same Likert scale throughout the questionnaire is crucial when running a factorial analysis.
- In rating questions (Likert scale, e.g., a little, a lot, too much or rarely, often, always), the number of grades must correspond to the respondent's discretion - usually a maximum of five.
- The polarization of answers should be avoided. For intance, not all answers with a negative connotation should be on the same side of a Likert scale, so that the respondents will not feel bad if their answers about their habits, way of life etc., are on the negative side of the spectrum. To achieve this, a reversal of positive and negatives answers should be used.
- Choosing between odd and even scale answers should be chosen depending on the seeking of polarization in the responses or the opportunity for neutral attitude towards a subject.

Questions and answers to be avoided

- *What would you do if…? Would you like …; What would you choose…?* In these questions the respondent is asked to answer hypothetical scenarios. As already noted, answering questions for which the respondent has no experience is useless, since they have very limited or no validity at all in the case of needfinding. This category also includes questions that may be asked to future users of a system about specific design solutions, before they try them (e.g., what kind of switch would you prefer?)

- *How often do you do…?* without specifying the time period we are referring to (e.g., day, week, month) In fact, the answers are not useful, since each respondent probably considers a different period of time. In addition, it should be noted that, even if we specify the time period, respondents often answer what they would like to do and not what they actually do. This is why we prefer asking questions that require recall from recent memory as stressed above (e.g., "How many times have you worked-out in the last week?").
- *Would you like to work with less effort?* Would you like to save time? Would you like to save money? Leading questions through the use of value-laden expressions or words like these direct the respondent to give us an obvious answer. The word "*cluttered*" in the Walmart example is an example of a badly connoted word which creates a leading effect on the question.
- *Do you go out often? How much do you like fruits?* Questions like these are quite vague and each respondent may perceive them differently. For example, the words "*often*", "*go out*" or even *"fruits*" in general are open to many interpretations and do not provide minimum and maximum values.
- *How many times have you caused an accident while driving?* Questions like this incriminate the respondent, putting him in a difficult position and the answer is very likely to be questionable in terms of honesty. In such cases and if the information is necessary, indirect questions should be preferred, from which the requested information will be deduced. For example, "*how many times have you had an accident while driving?*" and then "*describe to me how each accident happened*".
- *What is your age? What was the grade of your degree?* Questions like these are quite personal, and for some people not pleasant to answer honestly. It is thus recommended to ask the respondents to classify themselves in wider categories, for example in decades in the case of age.
- Questions that trigger prestige bias. This is the tendency for respondents to answer in a way to conform to social norms. For instance, in the question "*How many books have you read last year?*" someone who does not read books might feel the urge to provide a false answer, since reading books is considered to be an activity of high value in our society. The same question can be expressed as follows "*Did you have time to read a book this year*?" removing the responsibility for not reading from the respondents and thus probing them to answer more freely.
- Questionnaire transparency towards a certain goal by which the respondents feel affected, can also lead them to provide false answers that seem desirable or expected. For example, "*have you ever been fired*" in job application form may lead respondents to answer falsely.

Two basic principles that we should always strive for when conducting interviews or questionnaire surveys are:

1. Validity. The interview or the questionnaire should provide us with the information we really need and as close to the truth as possible.
2. Reliability. The data that will be collected and the results that will be extracted from their processing should be reproducible, if someone else repeats the same interview or questionnaire to the same population.

Overall obtaining reliable results on user needs from surveys can be challenging as these depend on the ability of respondents to clearly state their answers without being biased or affected by the survey process. That is the reason why we generally avoid relying exclusively on surveys, but instead we use them as supplementary methods.

4.4.3 Focus Groups

Meghan Ede, Professor and User Experience advisor in industry, once ran a series of several focus groups with IT System Administrators. When she first asked them what they did, they said things like: install, upgrade, oversee, and troubleshoot XX Operating System. Their work sounded very technical, as if they spent most of the day in front of a computer. When she next asked them to tell her about their last full day at work, starting with whether they had a cup of coffee in the morning, their answers were strikingly different. They all launched into stories about how they couldn't drink coffee because they would be accosted on the way to the coffee machine with requests for help. Most of them carried notebooks to record end-user problems. They talked about telling users over and over again how to do the same simple tasks, like change passwords, or screen backgrounds. They talked about training courses and surfing the web, being paged at home and reading technical books in bed. They read incessantly; newsgroups, bulletins, whatever they could find, in the hope of finding answers to problems not yet encountered. Professor Ede states that she could never have learned this in a usability study (which asks if a specific tool does its job well) or in a survey (few IT System Administrators realize how much time they waste in answering informal questions). A customer site visit would have taken days or weeks, not hours, and wouldn't have covered such a broad range of companies and positions.

Focus groups are group interviews involving a small number of people (usually 5-8) with common traits or experiences and a moderator / facilitator / researcher. The interview typically lasts about 2-3 hours, during which the moderator poses questions and maintains the group's focus. The participant's answers are recorded, sometimes by the moderator sometimes by an observer, and then, analyzed and reported at the end of the process. This method is preferred over questionnaires when the study concerns specific users (i.e., in a redesign of a corporate system interface) instead of a large audience with different levels of expertise and various needs.

Contrary to marketing-oriented focus groups, the participants in a need-finding focus group should not be asked to choose between alternative design solutions but instead to focus on their habits or problems with existing systems. The questions are usually open-ended and focus on understanding how typical or expert users behave or react in certain circumstances. As in the case of questionnaires and interviews, no hypothetical scenarios should be discussed but only real situations that have occurred in the past. The advantage of a focus group over individual interviews is that a participant's story might trigger the memory of the others who can contribute and light up other aspects of similar incidents and experiences. The downside is that in many cases participants need privacy to disclose certain experiences, being reserved to do so in front of colleagues or to a group of strangers. Especially when participants consist of people from different departments or different hierarchical levels of a company, they tend to conceal some aspects of work as done and resort instead to work as prescribed. So, although focus groups can be a powerful tool in the need-finding process, they should always be complemented by other information sources.

Some good practices and things to avoid during focus group sessions (adapted from the Interaction Design Foundation [2022]):

- The moderator should clearly explain the purpose of the group and what is expected of it.
- The moderator should try to establish a permissive environment in which everyone feels free to contribute.
- The moderator's job is to progress the discussion and to facilitate it and not to participate in the discussion itself.
- The moderator may probe for understanding if they feel that someone is on the verge of an important insight.
- The moderator should sum up important points at convenient moments and ensure that all participants have understood them.
- The moderator (with the observers) should lead a summary exercise at the end to summarize key themes, check for understanding and ask any questions that the observers feel would be useful.
- If any kind of recording is to be used, it should be explained in the introduction.

4.4.4 Sporadic or distant information gathering techniques

Let's consider that you are asked to design an IT application to assist the work of a winemaker. Such a project would require a meticulous need-finding study covering all the stages of winemaking. However, winemaking is a rather slow, event driven process, varying from season to season and the winemaker monitors and intervenes in the vineyard and the vinification process intermittently throughout the year. It is thus, next to impossible, time and cost wise, for a researcher to observe the whole winemaking process in the field. In addition, it is also quite unreliable to obtain such information through interviews if these are done out of context. Therefore, the aforementioned need-finding techniques would be either infeasible or unreliable. Fortunately, a number of reliable techniques have been developed specifically for activities that are either distant, sporadic and/or time consuming such as the monitoring of a medical treatment or the observation of a remote researcher.

Diary Studies

As Kim Salazar (2016) of the Nielsen Norman Group notes, a diary study is a research method used to collect qualitative data about user behaviors, activities, and experiences over time. In these studies, data is self-reported by participants longitudinally, that is, over an extended period of time that can range from few days to even a year or more. During the defined reporting period, study participants are asked to keep a diary and log specific information about activities being studied. To help participants remember to fill in their diary, they may be periodically prompted (i.e., through a notification received daily or at select times during the study period).

A diary study, unlike other common user-research methods, such as surveys and interviews, is considered to be a reliable need-finding technique due to the contextually appropriate time of information collection. Although time consuming and not being able to provide the richness and detail of field studies, diary studies are quite popular due to their minimum resource requirements.

In such studies the actual load is transferred to the participants that must be willing to cooperate. To achieve this, the most appropriate way is to build up internal motivation. In addition, study participants must have all the support they might need during the research, either by giving them full insight on the purposes and the importance of the study or by encouraging them and rewarding their achievements when they reach specific milestones. Giving periodic reminders to accomplish certain tasks can also be helpful when done with moderation so as not to cause annoyance. If the study requires heavy participant involvement, external motives can be considered such as monetary reward to compensate the time and effort given, but only as an additional incentive along with the above intrinsic motives.

To conduct a successful diary study Kim Salazar suggests the following good practices:

- Make sure the study is long enough to gather the information you need, but be cautious about designing a very lengthy study. If the study is too long, participants may become less engaged as the study progresses, which could result in less accurate data.
- Recruit dedicated users. Since diary studies require more involvement over a longer period of time, be extra prudent in the recruiting process. Let users know what is involved and expected of them up front. Ask screening questions that will help you gauge the level of commitment you will get from them during the study and be sure to confirm they will be available for the entire study period.
- Be on top of the data as it comes in. If you are getting data digitally or immediately as it comes in, evaluate it right away. This allows you to ask follow-up questions and prompt for additional detail as necessary, while the activity is still fresh in the minds of the participants.
- Conduct a pilot study. Diary studies can take quite a bit of time to plan and conduct, so it is helpful to conduct a short pilot study first. The pilot study does not need to be as long as the real one and it is not meant to garner data for analysis. Its purpose is to test your study design and related materials. Practice the process of briefing and debriefing pilot participants. Try out your logging materials to be sure they are understandable. Tweak your instructions and approach to ensure you get the data you need. Ask pilot participants for feedback about materials and the diary study experience and adjust accordingly.

Experience Sampling Method

As we saw, diary studies require significant commitment from the participants and this is their weak spot in terms of reliability, as the whole process depends on their persistent engagement. To overcome this weakness, a more efficient yet intrusive method used in certain contexts is experience sampling (ESM). The method, developed by Larson and Csikszentmihalyi (1983), is based on asking participants to report on their thoughts, feelings and behaviors on specific moments or occasions over time. Participants report on these, at the moment or shortly after some crucial events, after receiving an active reminder or notice to do so.

Increasingly, ESM is being used as a clinical monitoring tool in psychiatric and psychological treatments. In a study published in 2012, by T.J. Yun of the Georgia Institute of Technology, 30 pediatric patients with asthma received SMS text messages over a period of four months, with questions about their symptoms and with information about their disease. Questions like: "*In the past 4 weeks, did you have wheezing or difficulty breathing when exercising*?" and "I*n the past 4 weeks, did you miss days of school because of your asthma?*" were send and their responses were collected by the collaborating physicians. The patients decided what time of day they would like to receive the queries. When a child patient received a message, s/he could reply to the message by entering 'y'/'n' (yes / no) or 't'/'f' (true / false) (Figure 4.13). For knowledge questions, the SMS service sent the correct answer regardless of the user's response. The physician's dashboard was populated with their patients' responses and allowed the medical stuff to actively monitor the patients' status.

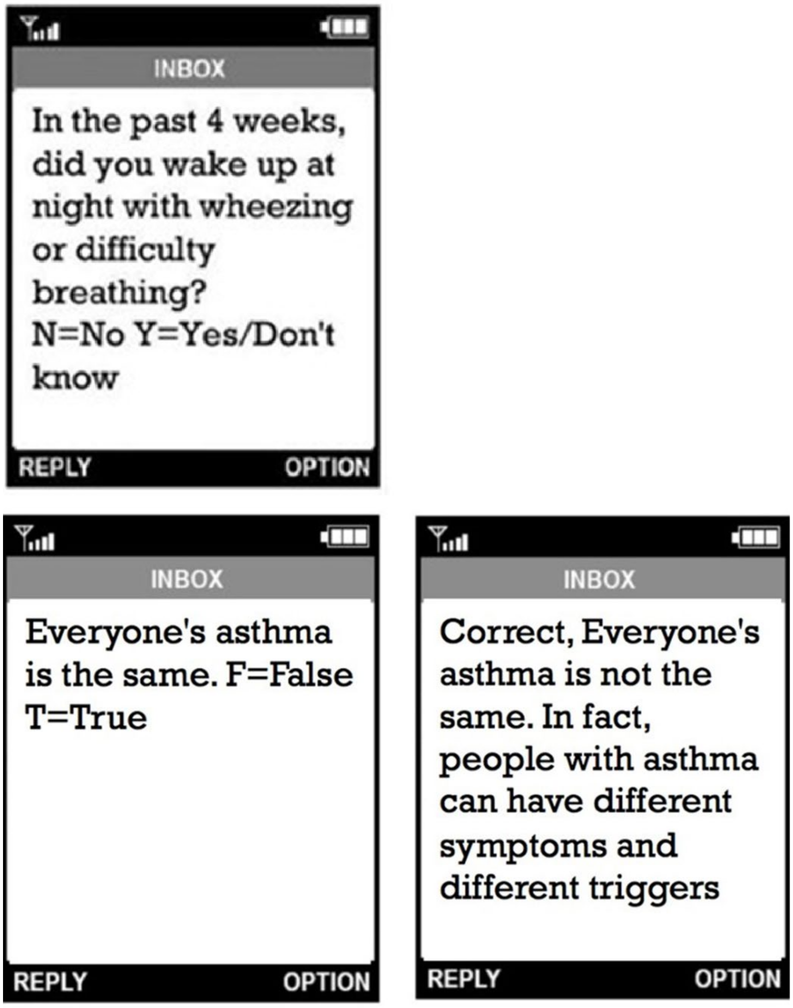


Figure 4.14. Some of the SMS received by patients with asthma during the study of the Georgia Institute of Technology.

Eventually, the patients who participated in the study, showed improved pulmonary function and a better understanding of their condition within four months, compared to other groups.

In the above case study, the ESM contributed to gathering reliable and direct information, from patients outside of hospitals. Contrary to a diary study, the procedure ran in real time so the researcher could instantly intervene if needed. The technological system that was built for that purpose was of course much more expensive than a simple diary, but in that case the benefits clearly outweighed the cost. Nowadays, we can think of many alternative technological possibilities to actively monitor participants on similar studies with reminders triggered automatically through sensors or that could offer much more flexibility and room for ad hoc interventions.

In summary, the Experience Sampling Method is a need-finding technique which:

- works through active reminders and notifications,
- is particularly well adapted for monitoring event driven behavior,
- can be manually or automatically triggered through the use of sensors,
- is very reliable because of its immediacy,
- might be annoying to participants, as it is not paced by them,
- needs significant recourses to design and implement.

4.4.5 On the choice of methods

As discussed earlier, not all methods give similar results, and this is mainly because they target to the different views of work as depicted in Figure 4.4. Log files, records, and regulatory directives, although explicit and objective, do not always capture the essence of the activity (i.e., work as done) since they tend to reflect a managerial view. On the other hand, the survey methods while purely "worker-centric" tend to give a theoretical view of how the work is done (i.e., work as espoused). Figure 4.14 depicts a rough mapping of the need-finding methods on the different views of work, which unveils why direct field observations are considered the most valuable source of data in activity analysis, but also how these methods complement each other in order to get a holistic insight of the work system.

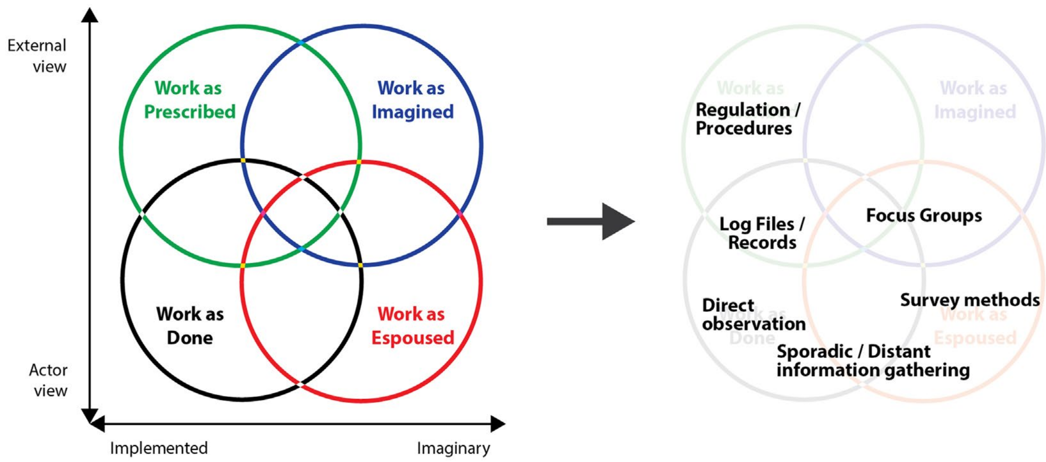


Figure 4.15. The mapping of the various need-finding methods on the different views of work.

Technically, all the methods and techniques presented above can be used exclusively or in conjunction to one-another, in order to obtain the best possible information depending on the particular study and available resources. Furthermore, each method or technique can be tweaked to better suit the needs of each particular study, while observing its limitations. Intricate knowledge on how people behave and act when being observed or questioned, will also help us devise new innovative ways of capturing their behavior and needs in the most ecological (i.e., natural) manner.

The responsibility, therefore, for the right choice of method(s) in each project lies with the researcher. Overall, when choosing methods, three main factors should be considered:

Focus: What kind of information are we seeking, e.g., objective, behavioral, attitudinal, affective?

Validity / Reliability: Which method provides the most valid and reliable information about what we want to learn.

Cost: What will be the cost of the method we choose in terms of time, effort and financial resources, to collect and process the information in question?

Balancing the last two factors is often a matter of tradeoff but depending on the available resources and the gut feeling of what would be promising to know, an experienced researcher can combine or adjust methods accordingly to succeed in getting the most critical and reliable information with the least possible resources.

References

Daniellou*, F., & Rabardel, P. (2005). Activity-oriented approaches to ergonomics: some traditions and communities. *Theoretical issues in Ergonomics science*, *6*(5), 353-357.

Garrigou, A., Daniellou, F., Carballeda, G., & Ruaud, S. (1995). Activity analysis in participatory design and analysis of participatory design activity. *International journal of industrial ergonomics*, *15*(5), 311-327.

Hucho, W. H. (2005). Der Luftwiderstand von Personenwagen*. Aerodynamik des Automobils: Strömungsmechanik, Wärmetechnik, Fahrdynamik, Komfort*, 157-284.

Hutchins, E. (1995). *Cognition in the Wild*. MIT press.

Hutchins, E., & Klausen, T. (1996). Distributed cognition in an airline cockpit. *Cognition and communication at work*, 15-34.

Interaction Design Foundation (2021) *How to Conduct Focus Groups.* Interaction Design Foundation https://www.interaction-design.org/literature/article/how-to-conduct-focus-groups

Larson, R., & Csikszentmihalyi, M. (1983). The Experience Sampling Method. *New Directions for Methodology of Social & Behavioral Science*, 15, 41–56.

Nathanael, D., Portouli, E., Gkikas, K., & Papakostopoulos, V. (2012). What does a motorcyclist look at while driving at urban arterials?. *Work*, 41(Supplement 1), 4900-4906.

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A. (2018). Naturalistic observation of interactions between car drivers and pedestrians in high density urban settings. *In Congress of the International Ergonomics Association* (pp. 389-397). Springer, Cham.

National Highway Traffic Safety Administration. (2014, March 28) *Tesla Motors, Inc Petition for Rulemaking*. Regulations.gov. https://www.regulations.gov/document/NHTSA-2010-0162-0254

Oppenheim, A.N. (2000). *Questionnaire Design, Interviewing and Attitude Measurement*. New York: Bloomsbury Academic.

Salazar, K., (2016, 5 June). *Diary Studies: Understanding Long-Term User Behavior and Experiences.* Nielsen Norman Group. https://www.nngroup.com/articles/diary-studies/

Shorrock, S., (2016, 5 Dec) *The Varieties of Human Work.* Humanistic Systems. https://humanisticsystems.com/2016/12/05/the-varieties-of-human-work/

Suchman, L. A., (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

Whalen, J., Whalen, M., & Henderson, K. (2002). Improvisational choreography in teleservice work. *The British Journal of Sociology*, *53*(2), 239-258.

Yun, T-J., Jeong, H.Y., Hill, T.D., Lesnick, B., Brown, R., Abowd, G.D., Arriaga. R.I., (2012). Using SMS to provide continuous assessment and improve health outcomes for children with asthma. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/2110363.2110432

Chapter 05:
User Requirements analysis/ Personas / Use Case scenarios

Chapter 05: User Requirements analysis/ Personas / Use Case scenarios

Chapter Summary

The chapter provides guidance on how to translate data from the need-finding process above and structure/ represent them into a cohesive format in tandem with technical constraints and process owner requirements. Techniques for abstraction and consolidation of requirements are presented through real world examples and case studies for class use. Next, an introduction to the use of Personas as a means to create empathy for the various end users and their usefulness on the building of use-case scenarios is presented. The chapter ends with guidance on developing use case scenarios to be used as reference at the latter stages of the design process.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of previous chapters.

5.1 Translating needs to requirements

The need-finding process discussed in Chapter 4 will produce a fair amount of diverse data, from pictures and field notes to eye tracking scan paths and open answers on conducted surveys. However, the scattered nature of such data makes that they may largely remain unused on the next phases of the design process if one fails to structure them into a coherent way. However, the whole essence of the user centered design process is to genuinely follow the need-finding outcomes and transform them into formal "User Requirements". User requirements are general statements often in the form of a "wish-list" or alternatively a "design brief" destined to drive main design choices. The transformation of user needs to requirements is the first step of the conceptual design phase. User requirements should be general enough as to not specify technical means of implementation; however, they should pragmatically take into account the technological possibilities available. In the present chapter we will look on how the results of the need-finding process can be structured and translated into such formal requirements and design parameters.

5.2 A case study: Automatic plant water timer

Consider a request for the design of an automatic plant water timer for domestic use. Such a timer usually consists of a solenoid valve, related electronic circuitry and a user interface. It is connected to a water supply and distributes water through a pipe system with adjustable nozzles providing each plant with the appropriate amount of water (Figure 5.1). The only a priori constraint imposed for the design is that the timer interface shall not include a digital screen but must consist only of physical knobs, sliders and buttons.

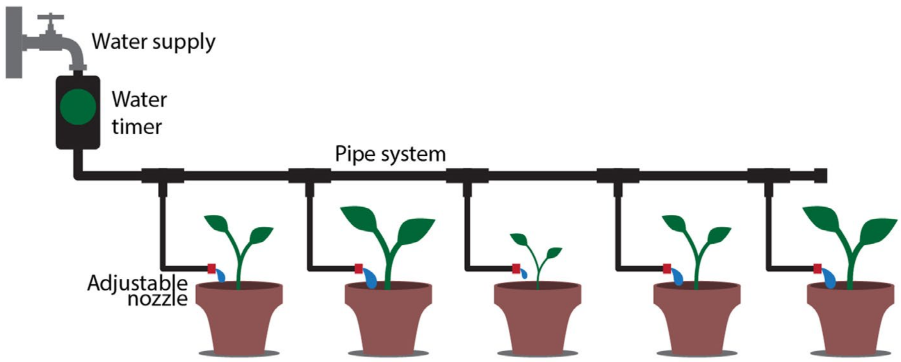


Figure 5.1. A typical plant watering system composed of water timer, piping and adjustable nozzles.

The needfinding of such a design request could include inquiries of several actual user setups of such systems, along with brief interviews with those users on what made them choose these setups, how do they change them throughout the year, or potential scenarios where current system interfaces do not cover their needs. We would also conduct interviews with domestic plant caretakers that manually water their plants so as to record their habits on watering and ideally an interview with an agronomist to understand the watering requirements of such plants. The results of this phase would be photos, perhaps videos and also, notes and written answers. These would form the map of the needs and habits of the users but also the plants' watering requirements.

From our system inspections and photos, we would be able to make an analysis on which automatic setups are most commonly selected while the interviews would reveal more about the habits of the users. For instance, having conducted such an analysis, we have found that the most common irrigating durations are between 2 and 8 minutes and the most common intervals between 12 hours and seven days. The agronomist may suggest that proper irrigating should be done at night or very early in the morning, at least for the hot seasons of the year. Some recorded habits of users would be their social gatherings at home mostly in the evenings and weekends, when they prefer to avoid getting their balcony or garden wet from watering, and also their need to temporarily bypass the program to water their plants at any given time. Also, in some cases we have seen users being unaware of the exact settings of their system concerning start time (e.g., "sometime between 5 and 6 in the morning, sometime after midnight etc.) which shows that setting the starting time with minute-level precision is not really crucial, while the duration of the watering must definitely be on a minute level scale. We have also seen users keeping notes on their system, concerning the chosen settings as the interface itself did not provide an indication of the chosen setup. Finally, in our quick analysis we have found a definite correlation between short intervals and brief watering time and long intervals with prolonged watering time, for example 2 minutes every 12 hours or 6 minutes every 72 hours in the same season of the year.

Data as above should be the starting point for specifying design parameters. How many controls should this device have? How many and precisely which available settings and with what level of granularity. How can we cater for the need to shut the system off temporarily or program it to start at a later time? Is there a way to choose or avoid certain days of a week without resorting to a digital screen? Note that at this phase we do not concern ourselves with specific physical characteristics of the device such as its resistance to outdoor conditions or its size but focus on the fundamental functional parameters i.e., the choices / possibilities for action that the device should offer to its users.

In order to define such functional parameters, we should first consider the plant irrigation process in its most basic form. The main idea of irrigating is the deposition of a certain amount of water on the soil of each plant. This can be achieved in two ways: by immediate deposition of this amount at once (as it is more or less done when watering manually with a watering can) or by gradual deposition of water through piping for certain duration. By controlling the duration, we control the amount of water the plants will receive but as we saw we also need to consider the frequency and start time. The duration and start time can be either defined by determining duration and choosing a start time by pressing start (e.g., irrigate for minutes 5, start time now or in X hours from now) or by determining chronological start and finish times (e.g., start: 23.30 and stop: 23.35), in which case the duration is determined indirectly.

Figure 5.2 presents the possible alternative options one can choose from to solve the problem of setting up such a device. Each option has pros and cons, and it is the designer's duty to decide which one s/he will implement. For example, setting precise start and finish times solves both the duration as well as the time of day of watering. However, as an option it has one important drawback: the need to set an internal clock when the devise is put into operation for the first time. This entails both extra steps from the user and more controls on the interface. Furthermore, from the need-finding process we know that the starting time need not be accurate to the minute (e.g., "somewhere between 5 and 6 in the morning", "sometime after midnight" would be acceptable as definitions).

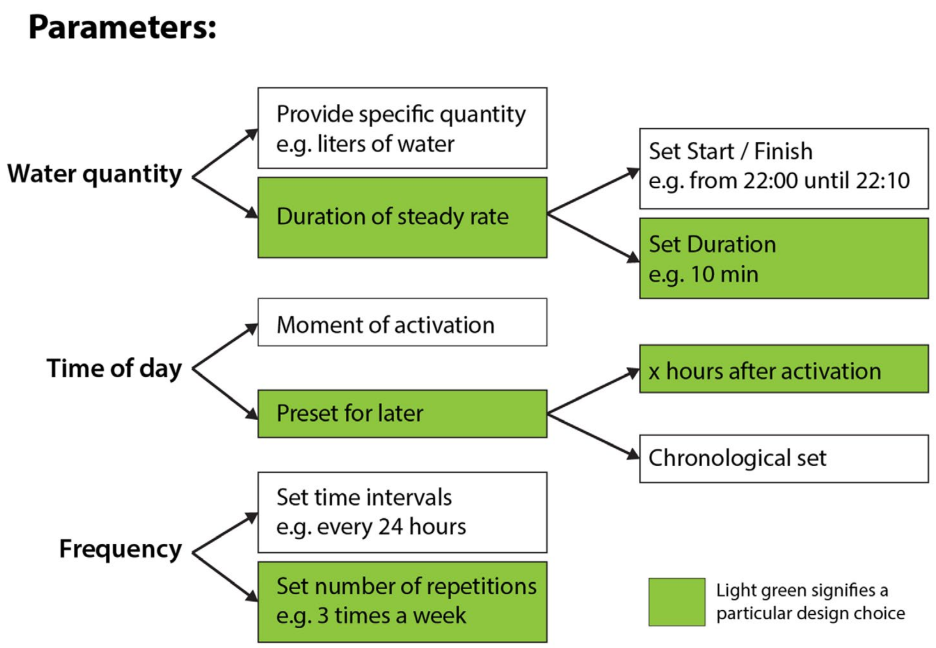


Figure 5.2. All the alternative configuration options for a watering system.

Figure 5.3 presents the basic user needs in the form of "formal" user requirements (i.e., water quantity, time of day, frequency) and their specification into design parameters. Design parameters, then, give rise to more detailed questions that the designer must answer.

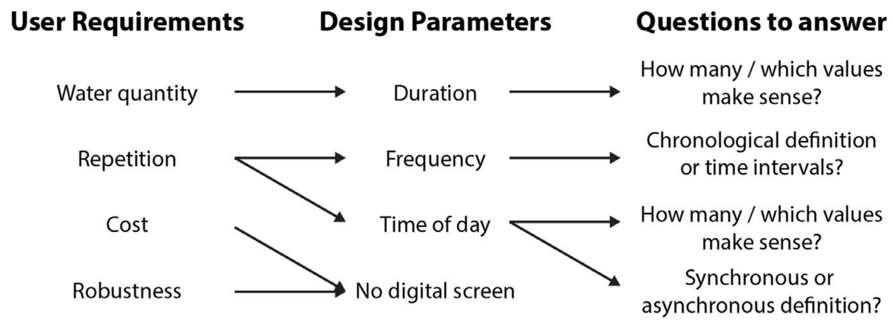


Figure 5.3. The user requirements forming design parameters which in turn are giving rise to questions that must be answered

One such question is: How many discrete options and what precise values should we specify for watering duration? It might seem like a trivial issue at first, but the core of a devise's usability rests on such decisions, e.g., on how well the watering options provided reflect user expectations and needs. An obvious way to address watering duration settings would be to select a fixed scaling with a constant interval (e.g., one or five minutes). However, if one looks deeper into the actual watering quantity options of such a solution, s/he would realize that each step increases the water quantity proportionally less and less. As illustrated in Figure 5.4, an increase in watering duration from 5 to 10 minutes results in 100% change in water quantity but as the numbers increase the percentages drop to a point that many options at the end of the scale become superfluous. This means that we add complexity and noise in the interface without real user benefit. On the other hand, the initial steps give 400% and 100% rise in water quantity respectively which are unacceptably large gaps according to user needs. How should we, then, proceed to balance the number and utility of options for the user?

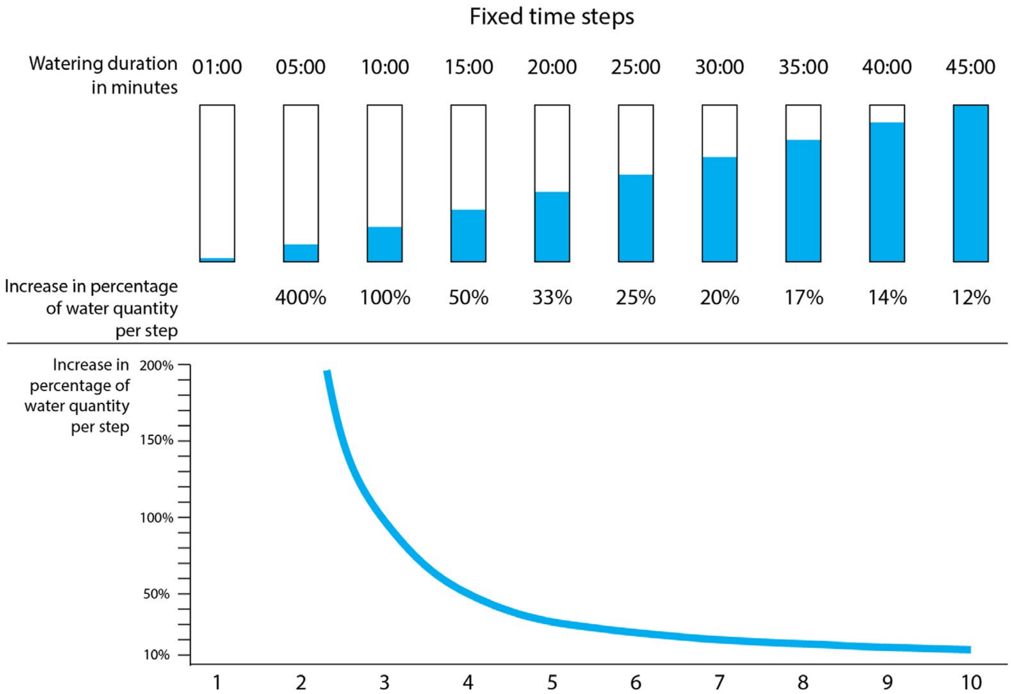


Figure 5.4. Equally increasing time steps on the duration of watering, will result in smaller and smaller increments in the deposited water quantity.

A sensible way to proceed is to provide a proportionally equal increase in water quantity at each step (a 50% increase being a fair requirement). To do that we would have to use the duration scale shown in Figure 5.5 below. Note that such a scaling is very precise and consistent to our initial requirement. On the downside however, the non-standard duration values (e.g., 2:15, 3:23, 5:04 minutes) are cumbersome to read and unintuitive for the user.

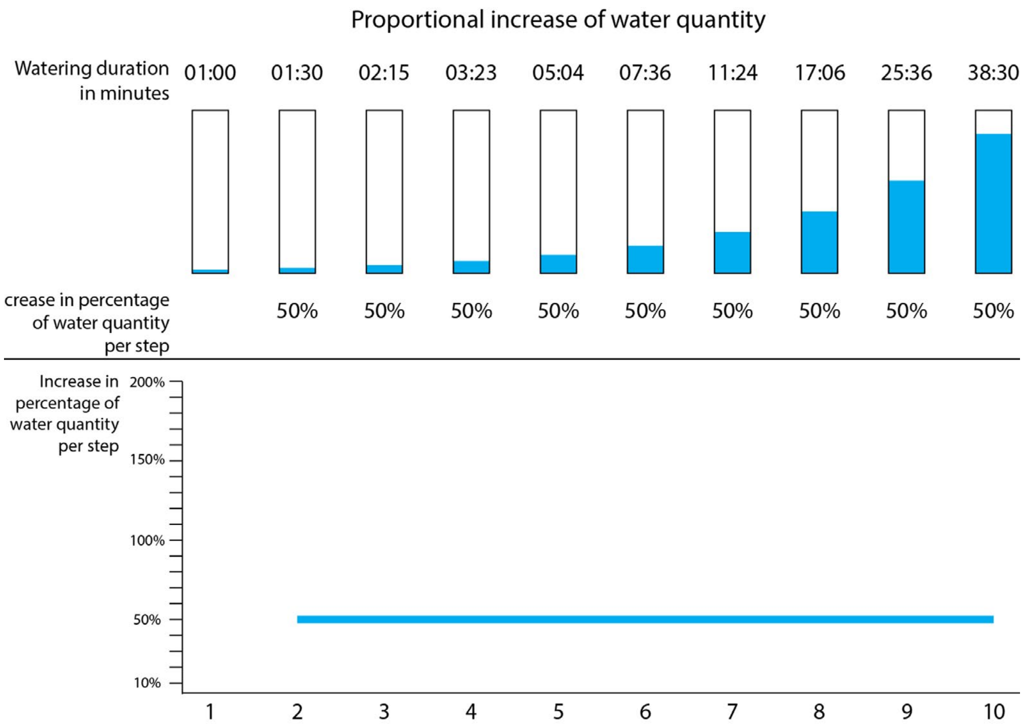


Figure 5.5. When selecting a stable increase of 50% in water quantity on each step, the resulting duration times look weird to the user.

Apparently, we have to steer a middle course between a consistent increase in water quantity and an unambiguous scale for the user. Figure 5.6 presents a reasonable tradeoff which is actually quite close to what most commercial watering systems use.

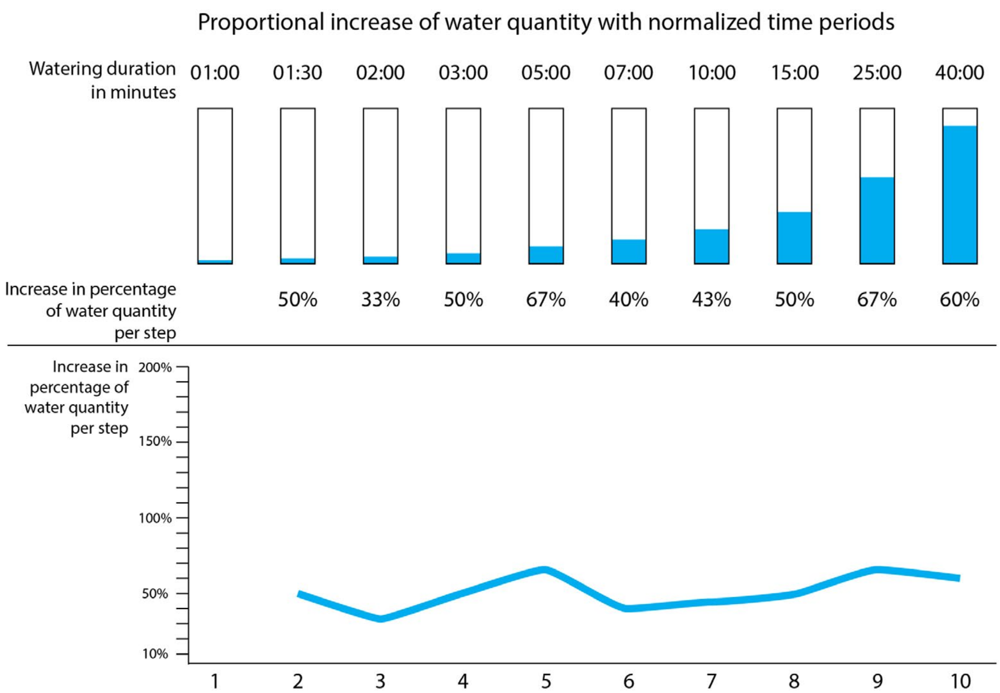


Figure 5.6. A good tradeoff between meaningful increase in water quantity and unambiguous time scale for the user.

Note that certain fluctuation in water quantity increments (from 33% to 67%) can be considered acceptable under the necessity for more clear-cut duration choices. Such tradeoffs are very frequent in a design process and the designer must be prepared to confront them adequately.

Regarding the repetition (frequency) of irrigation we should again consider the plants' biological requirements along with the users' social routines. While the daily and annual cycles are the only periodic phenomena that matter to the plants as biological organisms, the users are equally concerned for socially constructed cycles like weeks, months and holidays, as they schedule their routines also based on these social constructs. So, while a 12-hour and 24-hour periodicity is imperative, a weekly cycle is also essential to suit the users' needs. Lastly, the need to start watering at inconvenient social hours (e.g., late at night) makes it rather important to be able to program it in advance. For instance, if we want it to start watering at 4 am, it would be nice to set it in advance and not wake up to set it up for the first time (note that some commercial solutions do not offer this option for cost or simplicity reasons). The sequence of programming steps is not important as there is no real constraint in choosing the duration before the frequency or vice versa. So, design wise there is no need for a hard sequence imposition of the respective controls.

A multiple parameter system does not necessarily require the equivalent number of controls. Two or more parameters can be integrated in a single control, either with a multidimensional mode or simply by making some very effective compromises, as in some of the existing design solutions.

To illustrate the above ideas a number of different, commercially available, watering systems are presented below. Although they instantiate different solutions, each one targeting different users in terms of customizability and cost, all of them adhere to the design parameters and tradeoffs set in the requirements analysis above.

Figure 5.7. Two commercially successful models with almost identical interfaces, consisting of two rotary knobs and a push button.

The two systems in Figure 5.7 are almost identical in terms of interface design, incorporating three controls (two rotary knobs and a push button) with quite similar scales on each of them. The rotary knobs control the frequency and the duration while the push button can postpone the starting time by one hour with each press, thus achieving a basic asynchronous programming. The frequency scale includes not only the natural daytime cycles along with relevant fractions and multipliers but also a weekly cycle to satisfy the users' social routines. The duration scale has a similar approach to the one proposed above plus an option to bypass the programming which our need-finding analysis also identified as crucial for the users.

The design in Figure 5.8a has only two controls (a rotary knob and a push button) with which it accommodates all the functions described earlier. This is achieved by using the same knob for four distinct features by alternating its mode sequentially (current time -starting time - frequency - run time) with a push button. On the interface, a color code is used to indicate which scale around the rotary knob corresponds to each mode shown on the right. The downside of this otherwise clever design is that it lacks "memory in the world" as the user cannot later see which settings have been selected on each mode.

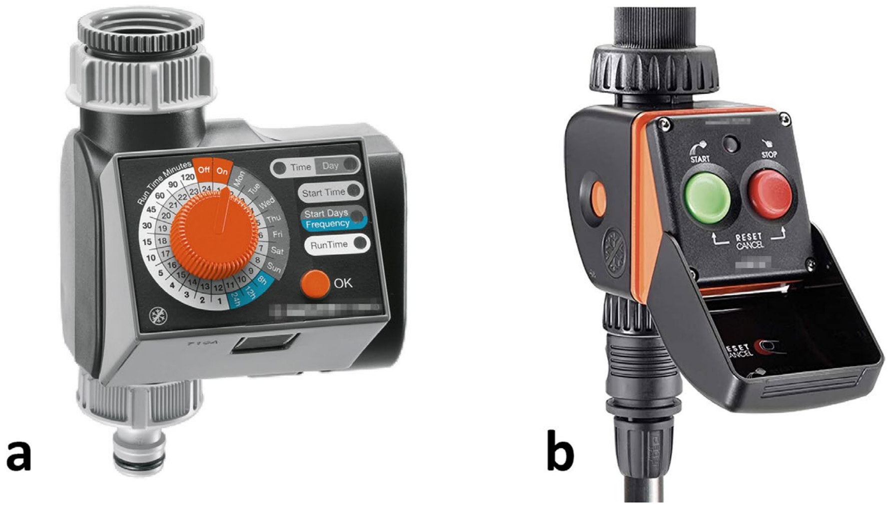


Figure 5.8.a. The left model incorporates a stepwise process based on a multimodal interface. b. The right interface takes a simpler approach where only the start and finish of the process can be set in real time.

Figure 5.8.b shows a very simple interface design consisting of only two buttons. The green button starts the watering while there red one stops it and activates a timer that will repeat the same watering duration every 24 hours. Its simplicity is reflected on its cost that makes it very competitive commercially, but there is also another benefit for the user: there is no need to decide in advance the duration that suits best the plant's needs. The user will simply start the procedure and terminate it when the water quantity seems adequate for the plants, thus having much more freedom in terms of watering duration. The compromises made in this design are that it only

offers a 24-hour periodicity and that the user cannot program it in advance but must be present at the exact starting / finishing time.

The design shown in Figure 5.9 has only one rotary knob but surprisingly covers most of the requirements discussed apart from the asynchronous programming. The idea behind this interface is that it utilizes the most plausible watering scenarios. For instance, the one-minute duration is seldom associated in real life with the 3 days frequency. Under this idea, only combinations that make sense are chosen and presented as predefined programs. The simplicity offered by this particular design can be achieved only after a very thorough need-finding phase and a subsequent requirement analysis as the one briefly presented above. A downside of the solution is that if the list of the programs is lost (a rather rare case since it is printed under the lid) the user has no clue of what the program numbers represent. Of course, one might say that it does not cover all the potential requirements compared to some of the previous ones, which is true. However, it is still commercially wise if you can satisfy 90% of the user population with a much simpler approach there is a very good chance that your product will become popular. The success of the design lies in making good compromises based on a solid analysis of the user needs.

Programme	Run time	Frequency
1	2 mins	12 h
2	5 mins	12 h
3	10 mins	12 h
4	1 mins	24 h
5	3 mins	24 h
6	5 mins	24 h
7	8 mins	24 h
8	10 mins	24 h
9	15 mins	24 h
10	30 mins	24 h
11	60 mins	24 h
12	5 mins	2 days
13	15 mins	2 days
14	30 mins	3 days
15	60 mins	7 days

Figure 5.9. An in-depth need-finding analysis can lead to a solution which combines only the meaningful combinations of duration and periodicity.

5.3 User Roles

In the automatic plant watering system case, we had a variety of users' needs and habits but all of them had the same goals and system configuration rights. If we were to examine a more complex system, such as an e-class or tax payment platform, we would see that their users may differ in terms of their goals, rights to post content, configure the system and supervise the actions of other users. In such systems we have at least two distinct user roles: the end-users and the system administrators.

Recognizing the different user roles involved in a system is crucial, as the needs and requirements of all roles must be taken into account. Accordingly, in a taxi service application we can recognize the role of the customer, the driver, the system coordinator etc. Each of these roles interacts with a dedicated interface and has different goals to fulfill. Therefore, the designer should identify and study all user roles throughout the need-finding and user requirements phases. In some cases, experienced and inexperienced users can also be accounted as different user roles, if access to certain system functions differs significantly between them.

5.4 Analyzing need-finding data: an example

Quantitative need-finding data analysis can be quite straight-forward. Either measurements or questionnaire data can be analyzed through direct statistical analysis (such as in data from closed-ended questions in questionnaires) or through content analysis (in open-ended questions). Purely qualitative need-finding methods, such as interviews and diary studies however, require a different approach. Each single evidence acquired from these methods may contain valuable information that should not be ignored even if there is no consistency or statistical significance throughout the set of data. Thematic Analysis is one of the most popular methods for analyzing such evidence.

Thematic Analysis is a method that involves identifying, analyzing, and reporting patterns (themes) within verbal unstructured data. It is an approach that relies on the researcher’s pre-existing thematic framework to guide the analysis. The patterns that emerge from the data are grouped together and used to develop a deeper understanding of the research topic.

Let’s look at an example from a study on ship pilotage, aiming to design a visual interface for navigation in closed waters. During the need-finding phase, several interviews were conducted with maritime pilots, to identify their cognitive tasks, difficulties, common errors and cues they use while piloting. (Parisi and Nathanael, 2019). Since the nature of the piloting task makes it quite difficult to observe critical maneuvers and conduct interviews in the field, a simulation of the task was utilized, and specific critical scenarios were examined. Vessel models from plexiglass and a printed nautical chart were used as low fidelity media to supp ort the interview while the whole simulation process was video-recorded. (Figure 5.10)

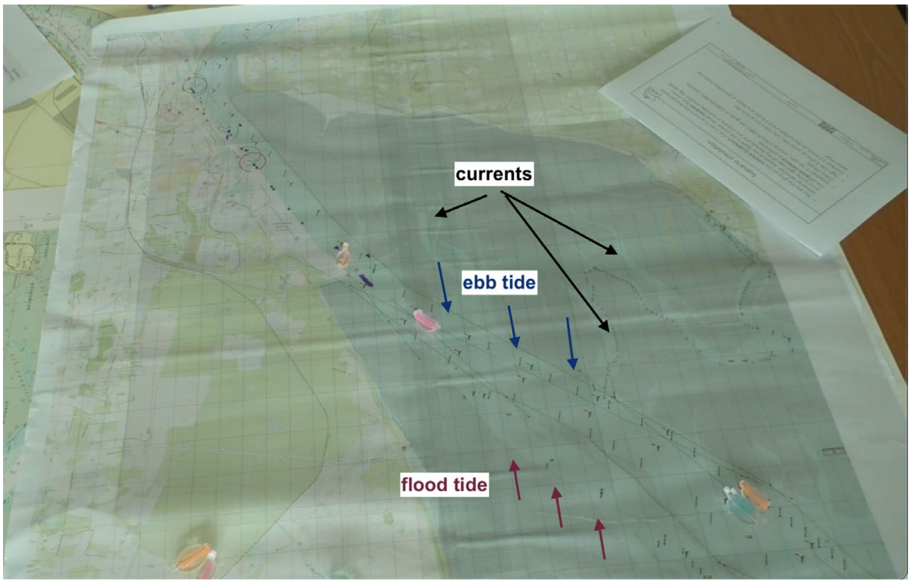


Figure 5.10. An annotated video capture image from a simulation interview on a scenario.

After the interviews have been transcribed, several phrases were tagged according to the thematic unit that they belong to (errors, difficulties etc.). Also, annotations were made on the transcripts, that show how the local environmental conditions (tide, current) affect the vessels that sail in the specific segment (Figure 5.11).

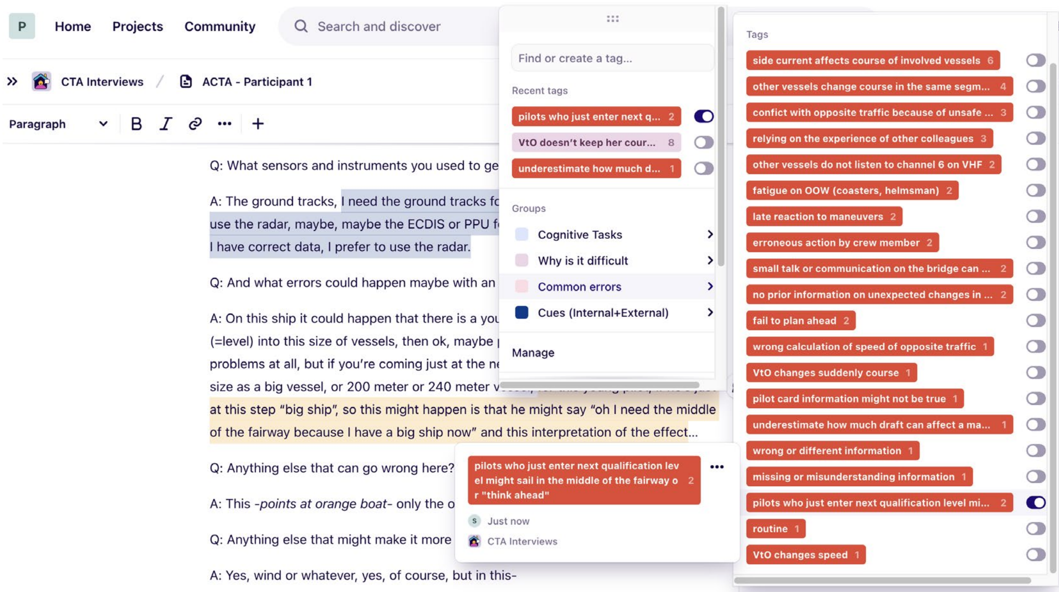


Figure 5.11. The tagging and annotation process of transcripts.

Finally, the themes that were used throughout the interviews (in this case: cognitive tasks, difficulties, errors, cues) are presented aggregated, giving a rich overview of the substantial parts of all interviews (Figure 5.12).

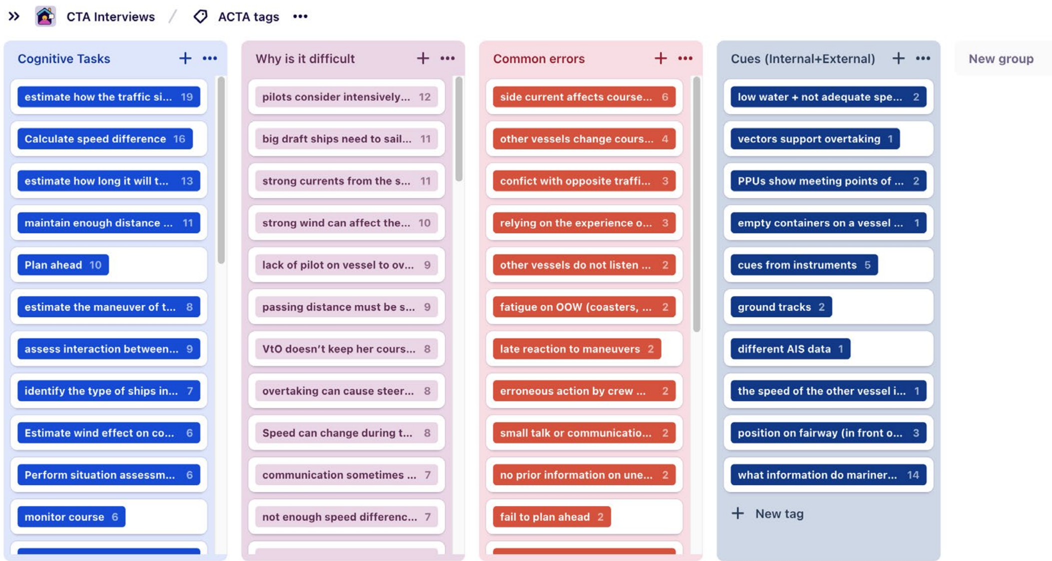


Figure 5.12. The resulting themes and tags after performing thematic analysis on five interviews.

Such an analysis can be done either manually by highlighting and annotating on paper or through specialized tools such as the one shown above. Overall, Thematic Analysis offers a systematic way of structuring specific pieces of valuable information scattered in unstructured verbal protocols. Such structured data can be directly used for the specification of formal user requirements.

5.5 Identifying user taxonomies: the Card Sorting method

In the income declaration form in the Greek tax system, there are two fields that are labeled “inputs” and “outputs”. Surprisingly these fields do not refer to the citizens’ incomes and expenses but to what the tax system considers as “inputs” i.e., the taxes paid by the citizen, and as “outputs” the eventual tax returns to the citizen. Indeed, it is quite common that the content of a form or a web page, is labeled based on what makes sense to the owner organization and not to its end users i.e., professional jargon that is almost incomprehensible to most users. Apart from the wording itself, another problem often faced when searching for information is the way this information is structured i.e., its “taxonomy”. The term taxonomy refers to how information is grouped, classified and labeled within an information environment. The resulting structure of that information environment based on a particular taxonomy is called the “information architecture”.

Here is an example from a bank website. Suppose you are a customer who wants to buy gold. Under which menu would you search when the available options are: “Deposits”, “Cards”, “Loans” “Investments”, “Insurance” and “Services”? Most users would choose “Investments”.

However, the Bank is resolute that this is a service, because, from their point of view, the "investments" category concerns internal bank products for which the bank has greater control over. Or what is more, suppose you want to send money to someone; would you identify your goal as a "Remittance"? Remittance might be a very common term in banking jargon but only few web users understand its meaning and match it to their goals (e.g., send money to someone).

To overcome such terminology issues, a technique called Card Sorting can be used. According to Katie Sherwin (2018) of Nielsen Norman Group, Card Sorting is a User Experience (UX) research method in which study participants group individual labels written on notecards according to criteria that make sense to them. This method uncovers how the target audience's domain knowledge is structured, and it serves to create an information architecture that matches users' expectations. In other words, it is a technique to gather user requirements concerning information architecture and labeling.

The basic set up of this technique consists of a pile of labeled cards ordered randomly. Participants must decide how these cards will be grouped depending on the given context. There are two main types of Card Sorting: the open and closed type.

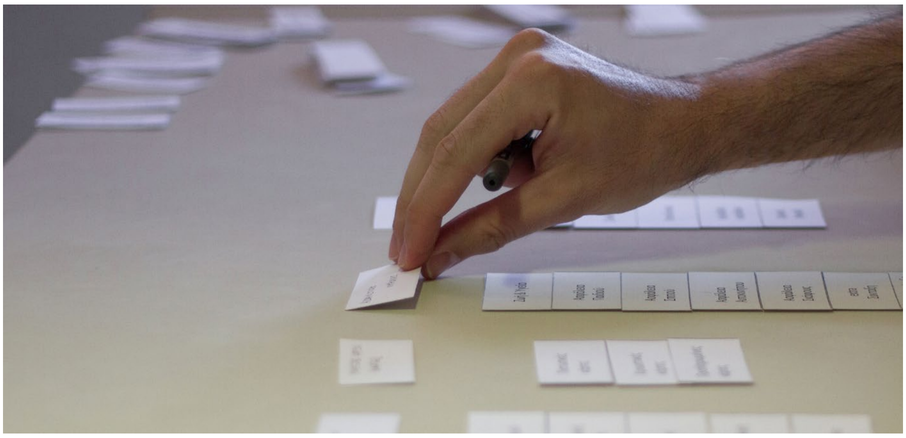


Figure 5.13. A participant giving a name to a set of cards, having previously grouped them together.

Open Card Sorting: Participants are asked to organize topics from content within a website or software into groups that make sense to them. Then, they are asked to name each of these groups in a way they feel that accurately describes their content (Figure 5.13). An open card sorting is preferred when we want to learn how users freely group content and the terms or labels, they would give to each group. Some users will make fewer groups with more content while others will form lot of groups containing fewer cards. In any case this is an in-depth exploratory process with a more demanding analysis.

Closed Card Sorting: Participants are asked to sort topics from content within a website into pre-defined categories. A closed card sort works best when you are working with a pre-defined set of categories, and you want to learn how users sort content items into each category. In that case the number of groups will be identical so the analysis of the results will be more straightforward.

You may also choose to try a combination of the two. You could conduct an open card sort first, to define the most promising content categories with one group of participants, and then, you could use a closed card sorting with a second participant group, to see how well the category labels work.

Card sorting will help you understand your users' expectations and understanding of your topics. It is often most useful once you have done some homework to find out about your users and understand your content. Knowing how your users group content in their mind can help you to:

- build the structure for your website,
- decide what to put on the homepage or toolbar,
- label categories and navigation

5.5.1 Good Practices for Card Sorting

A list of good practices for running Card Sorting tests is provided below (list compiled based on Usability.gov and Sherwin, K. (2018)):

- A group between 15 and 20 participants is considered enough for this technique. With more participants, you will get diminishing returns for each additional user; with fewer, you will not have enough data to reveal overlapping patterns.
- Try to limit the number of cards. It is tempting to want the participant to sort "ALL" of your content but be mindful of participant fatigue. We would recommend 30 to 40 at the absolute outside, especially for an open card sort.
- If possible, randomize the order of presentation, so that each piece of content has a chance to be sorted earlier in the session.
- Provide the participants with an estimate of how long the card sort will take, before beginning the session to help them better gauge the required time and effort.
- Consider the benefits of requiring participants to complete your sort. For an open sort, if possible, consider requiring them to sort the cards, but perhaps not to label them, since that might be the more challenging part of the task, providing you have limited your items as suggested above.
- Consider an open sort as "part one" and a closed sort as "part two" of your process. Part one allows you to learn what goes together, while part two allows you to really test out your labels to see if they are intuitive to your participants.

5.5.2 Analysis of Open Card Sorting

There are many Information Technology applications that can streamline the whole testing process and will also automatically analyze the results giving answers concerning statistical validity and correlation among the imported values. However, this technique can also be conducted with simple paper cards followed by data preparation and statistical analysis. In this case, once each participant has made his choices on grouping the labels, a photo of the results is taken, and the choices are inserted in a spreadsheet which produces a similarity matrix that will eventually show the correlation among pairs of labels (Figure 5.14).

Similarity Matrix

	Accounts	Consumer loans	Mortgage / Repair Loans	Investing	Mutual funds	Child Insurance	Life & Health	Extra Pension	Home Insurance	Car Insurance	Boat Insurance
Accounts		0%	0%	6%	6%	0%	0%	6%	0%	0%	0%
Consumer loans	0%		100%	12%	6%	0%	0%	0%	0%	0%	0%
Mortgage / Repair Loans	0%	100%		12%	6%	0%	0%	0%	0%	0%	0%
Investing	6%	12%	12%		88%	0%	0%	18%	0%	0%	0%
Mutual funds	6%	6%	6%	88%		0%	0%	12%	0%	0%	0%
Child Insurance	0%	0%	0%	0%	0%		100%	76%	100%	100%	100%
Life & Health	0%	0%	0%	0%	0%	100%		76%	100%	100%	100%
Extra Pension	6%	0%	0%	18%	12%	76%	76%		76%	76%	76%
Home Insurance	0%	0%	0%	0%	0%	100%	100%	76%		100%	100%
Car Insurance	0%	0%	0%	0%	0%	100%	100%	76%	100%		100%
Boat Insurance	0%	0%	0%	0%	0%	100%	100%	76%	100%	100%	
Time deposits	59%	0%	0%	35%	29%	0%	0%	18%	0%	0%	0%
Credit Cards	6%	29%	18%	0%	0%	6%	6%	6%	6%	6%	6%
Debit Cards	12%	24%	12%	0%	0%	0%	0%	0%	0%	0%	0%
Prepaid Cards	12%	29%	18%	0%	0%	0%	0%	0%	0%	0%	0%
Remittance	76%	6%	6%	0%	0%	0%	0%	6%	0%	0%	0%
Security Custody Services	18%	0%	0%	53%	59%	12%	12%	18%	6%	12%	12%
Lockers	41%	0%	0%	18%	18%	6%	6%	18%	6%	6%	6%
Stock Exchange Services	6%	6%	6%	88%	100%	0%	0%	12%	0%	0%	0%
PCP Support	0%	6%	6%	0%	0%	41%	41%	29%	41%	41%	41%
Gold Market	6%	6%	6%	94%	94%	0%	0%	18%	0%	0%	0%
Currency Exchange	47%	0%	0%	29%	35%	0%	0%	12%	0%	0%	0%

	Time deposits	Credit Cards	Debit Cards	Prepaid Cards	Remittance	Security Custody Services	Lockers	Stock Exchange Services	PCP Support	Gold Market	Currency Exchange
Accounts	59%	6%	12%	12%	76%	18%	41%	6%	0%	6%	47%
Consumer loans	0%	29%	24%	29%	6%	0%	0%	6%	6%	6%	0%
Mortgage / Repair Loans	0%	18%	12%	18%	6%	0%	0%	6%	6%	6%	0%
Investing	35%	0%	0%	0%	0%	53%	18%	88%	0%	94%	29%
Mutual funds	29%	0%	0%	0%	0%	59%	18%	100%	0%	94%	35%
Child Insurance	0%	6%	0%	0%	0%	12%	6%	0%	41%	0%	0%
Life & Health	0%	6%	0%	0%	0%	12%	6%	0%	41%	0%	0%
Extra Pension	18%	6%	0%	0%	6%	18%	18%	12%	29%	18%	12%
Home Insurance	0%	6%	0%	0%	0%	6%	6%	0%	41%	0%	0%
Car Insurance	0%	6%	0%	0%	0%	12%	6%	0%	41%	0%	0%
Boat Insurance	0%	6%	0%	0%	0%	12%	6%	0%	41%	0%	0%
Time deposits		0%	12%	6%	59%	29%	53%	29%	0%	35%	53%
Credit Cards	0%		88%	94%	12%	6%	6%	0%	35%	0%	0%
Debit Cards	12%	88%		94%	18%	0%	6%	0%	29%	0%	6%
Prepaid Cards	6%	94%	94%		18%	0%	0%	0%	29%	0%	6%
Remittance	59%	12%	18%	18%		6%	41%	0%	6%	0%	53%
Security Custody Services	29%	6%	0%	0%	6%		53%	59%	12%	59%	29%
Lockers	53%	6%	6%	0%	41%	53%		18%	6%	18%	35%
Stock Exchange Services	29%	0%	0%	0%	0%	59%	18%		0%	94%	35%
PCP Support	0%	35%	29%	29%	6%	12%	6%	0%		0%	0%
Gold Market	35%	0%	0%	0%	0%	59%	18%	94%	0%		29%
Currency Exchange	53%	0%	6%	6%	53%	29%	35%	35%	0%	29%	

Figure 5.14. A Similarity Matrix shows the percentage of correlation among entities as defined by the participants of the card sorting. A percentage of 100% shows that all participants agree on the correlation of a specific pair.

A hierarchical cluster analysis can, then, show the stronger groups, and the designer will be able to choose how many groups would be enough out of a correlation tree diagram (Figure 5.15). Since the participants are naming their groups, an extra analysis must be made concerning the labeling of each group in a qualitative manner.

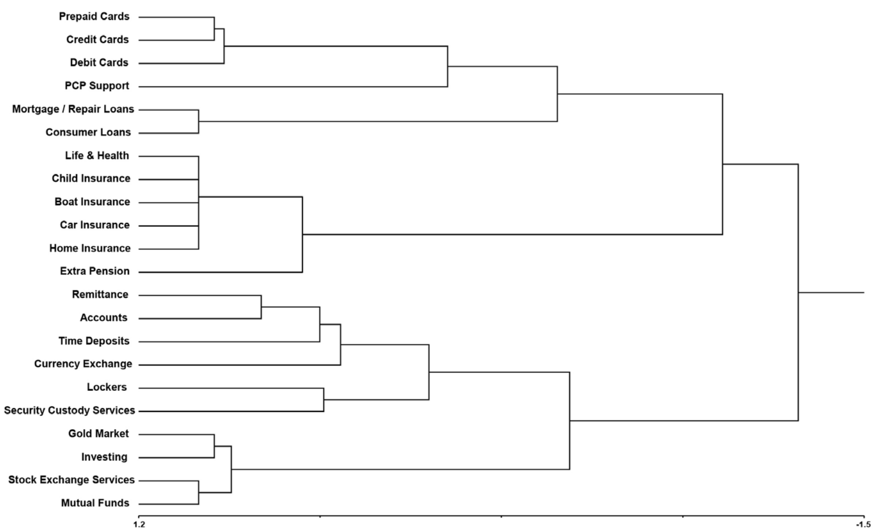


Figure 5.15. A hierarchical cluster diagram grouping entities at various levels of detail.

5.5.3 Analysis of Closed Card Sorting

In closed Card Sorting the analysis consists in grouping the cards according to the frequency of their occurrence in a certain category (Figure 5.16). Also, we can propose a headings' hierarchy based on the ranking order, selected by each participant and finally a hierarchy of the cards inside each grouping.

	Deposit Products	Banking Services	Cards	Loans	Investment Products	Insurance Products
Time Deposits	16	1	0	0	0	0
Accounts	12	5	0	0	0	0
Remittance	2	15	0	0	0	0
Currency Exchange	0	15	0	0	2	0
Lockers	3	14	0	0	0	0
Θεματοφυλακή Τίτλων	1	11	0	0	4	1
Security Custody Services	0	7	5	0	0	5
Credit Cards	0	0	17	0	0	0
Debit Cards	0	0	17	0	0	0
Prepaid Cards	0	0	17	0	0	0
Consumer Loans	0	0	0	17	0	0
Mortgage / Repair Loans	0	0	0	17	0	0
Investing	0	0	0	2	15	0
Mutual Funds	2	1	0	0	14	0
Gold Market	1	2	0	0	14	0
Stock Exchange Services	1	4	0	0	12	0
Child Insurance	0	0	0	0	0	17
Life & Health	0	0	0	0	0	17
Home Insurance	0	0	0	0	0	17
Car Insurance	0	0	0	0	0	17
Boat Insurance	0	0	0	0	0	17
Extra Pension	0	0	0	0	4	13

Figure 5.16. On the left column of the table, we insert all the entities that must be grouped, while on the upper line we place all the predefined category names. The numbers show the occurrence of each entity in a certain category according to the participants' answers.

It is not uncommon for this method to give results with small differences in the frequency of occurrence in a category. For example, the appearance of the same card in three different menus with percentages of 41%, 29% and 29% as shown in Figure 5.12, does not give a statistically significant lead to the first, resulting in either the need to increase the number of participants or better off try an alternative name for the card which may solve the ambiguity.

There are also other variants of this technique like the Reverse and the Modified-Delphi Card Sorting which can be used depending on the request, but the main idea remains the same: Uncover how the users' mental models are structured concerning taxonomy, and thus help us build an information architecture that matches the users' expectations.

5.6 Conceptualizing Model Users: The Personas

Personas are fictional characters, based on data that we have collected during the need-finding process and build in a way to represent commonalities among certain groups of users. While more popular in the marketing sector, personas are also useful to UX designers when building scenarios or thinking of alternative routes on a blueprint, as they provide coherent behaviors, outlooks, and potential objections of people matching a given persona. While not widely accepted as an essential tool and sometimes even criticized for promoting social stereotypes and reductive or biased user behavior, they might prove useful if used properly, acknowledging their pros and cons.

Personas help designers get out of their skin, and into the shoes of an archetypical user, in order to imagine their potential behaviors, goals and attitudes. Most often, personas are used to predict the behavior of these archetypical users on certain key tasks. For instance, a key task on an e-shop can be considered either the addition of items in the shopping basket or the actual buying of a product. Depending on the task, the UX designer tries to improve the experience of each fictional character (persona) by imagining what can go wrong in terms of the available options, the wording or even the symbols used at each screen that may push the user out of the website.

The characteristics of each persona must be tied back to real data collected from our prior research and observations of real users (Bendeler, K., 2022). After that, the most important of these are written down on a template along with some demographics (e.g., age, education, occupation), the goals and tasks that they are trying to complete using our product, their physical, social, and technological environment and also some personality indicators. Each persona must represent a real group of potential users and most of the times a total of 3 to 5 such personas are enough to describe most of the users.

An example of a persona description is shown in Figure 5.17.

Persona:	The senior manager
Photo:	
Fictional name:	Chris Brown
Job title	Senior Health and Safety manager
Demographics:	• 46 years old • Married • Father of two children • Has a Master's degree in Mechanical Engineering.
Goals and tasks:	He is focused, goal-oriented within a strong leadership role. One of his concerns is to predict and avoid potential hazards. Spends his work time: • Requesting and reviewing safety reports, • Supervising staff efforts in health and safety
Environment:	He is comfortable using a computer and refers to himself as an excellent Internet user. He uses email extensively and uses the web about 4 hours during his workday.
Hobbies:	Bicycling, woodworking
Quote:	"Predict soon, act immediately"

Figure 5.17. A typical Persona Card.

Based on such persona cards, designers try to form an objective ground for imagining how this type of user would interact with a system, what could frustrate her/him while trying to accomplish a task, and what s/he would find helpful. In order to fully utilize the possibilities of a persona we usually build use case scenarios that we, then, run with each of these personas and record the different behaviors that can occur.

5.7 Use case scenarios

Use case scenarios are a well-known user requirements analysis and specification technique, typically used in the design process of interactive systems that shifts the focus of the design from defining system operations, to describing how people use a current or future system to accomplish their goals (Rosson & Carroll 2002). By creating use case scenarios, the designer can obtain a better understanding of the users' needs through a concrete description of certain situations that might arise.

Scenarios can be classified in three distinct types depending on the design phase each type is used. At the user requirements phase scenarios are used to help the designers shape the problem space of a given design problem. These are called "problem scenarios" (Rosson & Carroll 2009) and include situations that are either not adequately covered by the current technological system or simply describe user habits and practices on similar tasks today. The problem scenarios are formed after a thorough need-finding work, when the designer feels confident in his grasp of user habits and needs. At the ideation phase, scenarios help designers to imagine possible future situations and to communicate those situations amongst the stakeholders of a new system. In these scenarios, personas are used to facilitate the process and get more fruitful results. Finally, user scenarios are also employed as scripts in in user tests in the system development phase.

To create a scenario the designer must describe a certain situation where a user (persona) will have to interact with the system within a certain context. Details such as the hour of the day, the place and specific user goals are considered critical so that all parties get a good grasp of the context. From our experience, writing a helpful scenario is quite demanding and lies mostly on the designers' experience of selecting the right blend of elements to get the most out of it. To make this process easier for the inexperienced designer, we have identified some aspects that make a scenario useful as well as aspects to avoid.

5.7.1 What makes a good scenario

Some useful guidelines to keep in mind when writing a scenario are the following (Gkikas et al. 2017):

- Describe situations not very obvious but nevertheless recurring amongst users. Typical scenarios are often ineffective because of their obviousness and thus their poor resolution in terms of user needs, while non-critical ones would translate in rather marginal and uninteresting findings. A number of elements that make an unusual but critical scenario are: (i) a situation which might be infrequent but absolutely necessary (e.g., "In a ferry booking website the option to send my car with the ferry but with no accompanying passenger"); (ii) a situation that affects many users under uncommon but plausible circumstances (e.g., "On their arrival at the port the passengers that bought a ferry ticket from the web are informed that the boat trip had been canceled due to bad weather"); (iii) misuses or unforeseen uses of the current system (e.g., "The users of the current bus service sometimes pass over their tickets to another user at the end of their journey when these are not overdue")

- Include two or more interacting "elements". Such elements may include specific user attributes, task type and an environmental disturbance or influence. By introducing more than one element in a scenario, not only do we get insight through their consideration one by one, but we also examine their interrelations, getting, in this way, a much richer picture of the problem space. For example, introducing a reference to a specific detail, e.g., "rainy day", the parameter "weather" is enriching the use-case with all its possible variations. Note that, although the problem scenarios' objective is to open up the problem space, being generic in the wording of scenarios does not help in this direction. On the contrary, being specific on more than one elements (like the identity of the user or the actual time that the incident takes place), not only creates more empathy towards the user, facilitating the dramatization of the story, but also enlarges the scenario coverage by inserting some differentiating parameters.
- Write scenarios through the eyes of a user group unfamiliar to the designers themselves. This "stepping in other people's shoes" is considered valuable in terms of widening the problem space. It is critical for the designer to be able to identify all the different users of the system. When a designer is studying a system which is familiar to him, there is the risk of overrating the situations already known to her/him, as s/he is biased from his own experience.

5.7.2 What to avoid in scenarios

Some guidelines on characteristics to avoid when writing a scenario are the following (Gkikas et al. 2017):

- Scenarios direct implying specific solutions. For example, in the scenario: "A tourist from the U.S.A needs to see the price of his ticket in USD" The resulting requirement is that "The system should give the option to convert the price in different currencies". This is an example of "solution-first" approach that is transformed in a scenario, so that it will arise later on, as a validated solution.
- Writing two or more scenarios which, in fact, describe the same use case by slightly changing one element. It might seem

as an effort to include all possible situations but, actually, it is of little interest to differentiate, for example, users eligible for discount by just changing their identity between "students" and "unemployed".

- Depicting a trivial situation directly deriving from the design brief. This usually happens because of the misconception that the larger the number of use cases the largest the area of the design space covered. However, we can safely presume that if one has thought of complicated use cases, they would also have catered for the trivial ones on the way. So, it seems rather uninteresting to describe the commonest so as to be safe in numbers.
- Depicting some extremely marginal scenarios (i.e., infrequent and/or irrelevant). Such scenarios might mistakenly divert the design process towards rarely arising situations, missing, in this way, other more core needs of the users.

Overall, the above guidelines should act as reminders of the integrative role of the designer and help prevent sterile and formalistic scenario writing. Typically, once an initial set of scenarios are first written down, designers need to review them as a whole and to re-express them in a way that enhances their complementarity and avoids overlapping.

References

Bendeler, K. (2022, Sep 2). *5 Steps To Build the Perfect Persona*. Medium.com https://medium.com/analysts-corner/5-steps-to-build-the-perfect-persona-df2e6c49d3be

Gkikas, K., Nathanael, D., Marmaras, N. (2017). Challenges faced when teaching how to write a user scenario. *Proceedings of the European Conference on Cognitive Ergonomics (ECCE '17).* Association for Computing Machinery, New York, NY, USA, 170–175. https://doi.org/10.1145/3121283.3121309

Parisi, S., & Nathanael, D. (2019). Adapting Applied Cognitive Task Analysis to identify cognitive challenges in sea pilotage. *Proceedings of the 31st European Conference on Cognitive Ergonomics* (pp. 69-74).

Rosson, M. B., & Carroll, J. M. (2009). Scenario-based design. *Human-computer interaction* (pp. 161-180). CRC Press.

Rosson, M.B. & Carroll, J.M. (2002) *Usability Engineering: Scenario-Based Development of Human-Computer Interaction.* Morgan Kaufmann Publishers, Burlington.

Sherwin, K. (2018, March 18). *Card Sorting: Uncover Users' Mental Models for Better Information Architecture*. Nielsen Norman Group. https://www.nngroup.com/articles/card-sorting-definition/

US Department of Health and Human Services*.* (accessed 2022, December 20). *Card Sorting.* Usability.gov https://www.usability.gov/how-to-and-tools/methods/card-sorting.html

Chapter 06: Conceptual design

Chapter 06: Conceptual design

Chapter Summary

The chapter introduces the reader to the actual design process beginning with conceptual design where tradeoffs among end-user requirements, process owner requirements and technical constraints start forming alternative solutions for the overall system. A number of representation tools are presented, which have been proved useful for conceptualizing i.e., abstracting and framing the design space. Specific tools include (i) Hierarchical Task Analysis that decomposes tasks to sub-tasks; (ii) State transition diagrams that describe system behavior, (iii) Flowcharts that represent the algorithm behind a user interface, (iv) Experience Blueprints that depict how a service is experienced by a user and, (v) User Journey Maps as a visual representation of the potential user's experience. Ideation, i.e., generation of ideas, per se is only tackled briefly at the end of the chapter, as it is considered beyond the scope of this book.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of previous chapters.

6.1 Creativity beyond method

The conceptual design phase is by far the most creative part in any User Centered Design process. There are neither well-defined methodological steps to follow nor strict rules to obey. Depending on the project and the preferences of each design team a variety of methods and tools may be employed. Most designers tend to adapt existing methods or invent their own to help them be creative, but at the same time organized and efficient. The selection of methods presented below is only indicative; their purpose being mostly to open up the designers' horizon in devising their own original methods or adaptations of existing ones according to each project needs.

6.2 Functional decomposition

If you have ever used a two-knob mixer tap in a shower (Figure 6.1), you would have found yourself constantly adjusting both knobs to achieve the desired flow and temperature. In fact, each time you want to adjust one of the two, the other is also affected (i.e., to increase the flow, you have to adjust both hot and cold knobs so as to keep the same temperature and vice-versa). If we analyze the user interaction with this interface, we will recognize four specific user goals that are realized through combinations or four possible physical user actions, i.e., Rotate Clockwise (CW) or Counterclockwise (CCW) the Cold knob, Rotate CW or CCW the Hot knob. These are shown at the bottom part of the diagram in Figure 6.1.

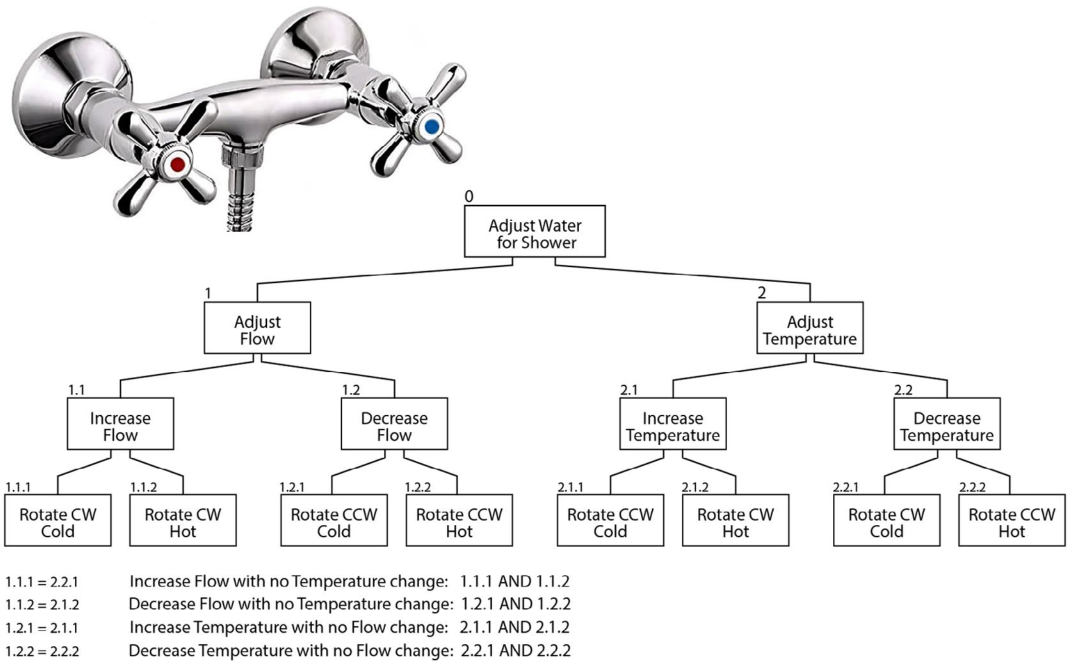


Figure 6.1. Task diagram for a two-knob mixer tap.

To build such a diagram we must first write-down all trivial user goals (or intentions) and then play the simple mind game of "asking HOW and WHY". By asking HOW we can trace the physical actions necessary for the accomplishment of each user goal. By reflecting on WHY we are progressively adding levels upwards, communicating in more abstract terms the reason we perform these goals, ending-up to our general goal, i.e., the adjustment of the water to take a shower. Tracing the diagram downwards, we will now see that our main goal is decomposed into alternative or complementary sub-goals, and these need more than one physical action to be realized.

Now let's examine the following diagram (Figure 6.2), which decomposes the same general task but, in this case, it involves using a single lever bath mixer tap.

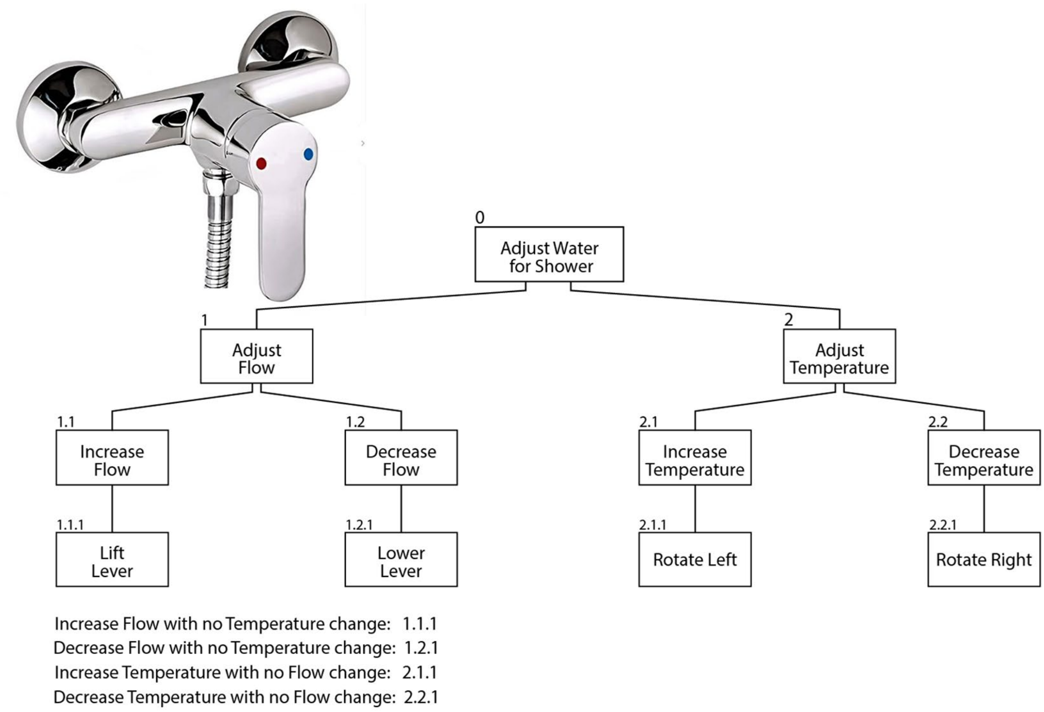


Figure 6.2. Task diagram for single lever mixer tap,

Note that with the single lever interface design, to accomplish his trivial goals the user needs to perform one direct action instead of two. In addition, the actions required for adjusting the flow are semantically separated from the ones that adjust the temperature. Hence, the popularity of single lever mixer tabs in contemporary homes. The type of analysis and accompanying diagrams used above to decompose the tasks and subtasks of the "adjust water for shower" process is called Hierarchical Task Analysis and will be demonstrated in more detail later in this chapter.

Consider now that you are asked to design a novel interface of an automated fuel station. The basic requirements for the design are that the fuel station must operate without the presence of an employee, and provide three types of fuel (Regular Gasoline, Super Gasoline and Diesel). Such a project obviously requires a thorough need-finding phase consisting of field study on several fuel stations, interviews with both employees and drivers, and analysis of all the recent receipts and invoices. Nevertheless, let's consider that the need-finding process is already done and now you need to focus on data interpretation, and then on specifying conceptual design alternatives.

Specifically, consider that during the needfinding, you have collected receipts with the following payments:

- 5% 5 euros worth of Gasoline
- 10% 10 euros worth of Gasoline
- 30% 20 euros worth of either Gasoline or Diesel
- 10% 40 euros worth of either Gasoline or Diesel
- 20% 50 euros worth of either Gasoline or Diesel
- 5% rounded amounts over 50 euros mostly for Diesel
- 20% receipts with unrounded amounts of payments for either Gasoline or Diesel.

After interviewing the employees of these fuel stations, you have learned that these unrounded amounts reflect either the need for filling-up the fuel tank or requests for specific amounts of liters (mostly coming from professional truck drivers or from clients getting fuel in canisters). On the other hand, the small amounts (5-10 euros) are coming from motorcyclists, while the rounded amounts over 50 euros mostly correspond to professional drivers. An interesting insight that derives from the above is that the transaction sequence may vary, depending on the request. For instance, whereas in most cases the price is set in advance, there are cases where the price cannot be calculated before delivering the fuel. Note that in conventional fuel stations all these cases are processed in the same sequence: first comes the fueling step and then the payment, since it is considered socially unacceptable to fuel up and leave without paying in the presence of an employee.

However, it is very unlikely that you would implement the same sequence in an automatic system, since the lack of human presence makes the system more vulnerable to malicious use. The request for a specific amount of fuel in liters, can easily be solved by converting the quantity demanded into a specific monetary amount, so that this amount can be deposited in advance. The tank filling-up, however, cannot be paid in advance since no a priori cost estimate can be made (Figure 6.3). That could lead the designer to envision a solution where the car is somehow locked in place after been filled up with fuel and would not be released until the payment is completed. However, that raises various safety and legal issues, not to mention that the fueling station itself would be put temporarily out of operation, a case with no actual benefit for anyone involved.

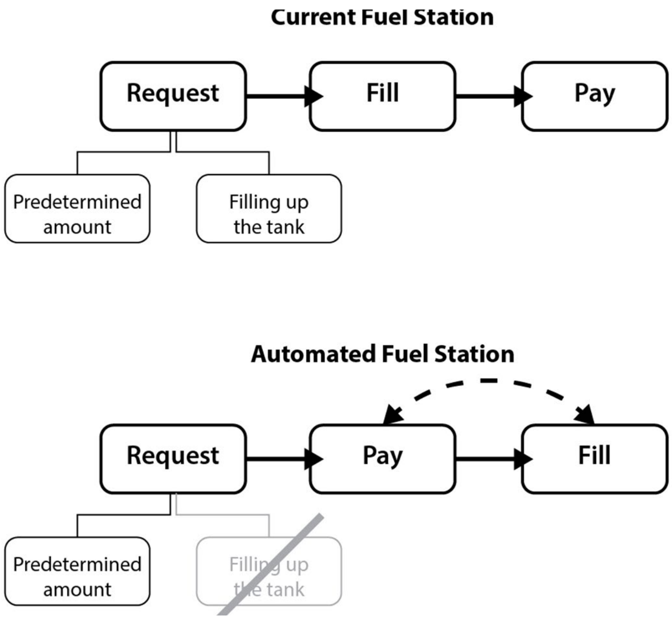


Figure 6.3. When the process of payment precedes the filling process, the option of filling the tank cannot be fulfilled as the amount cannot be predefined.

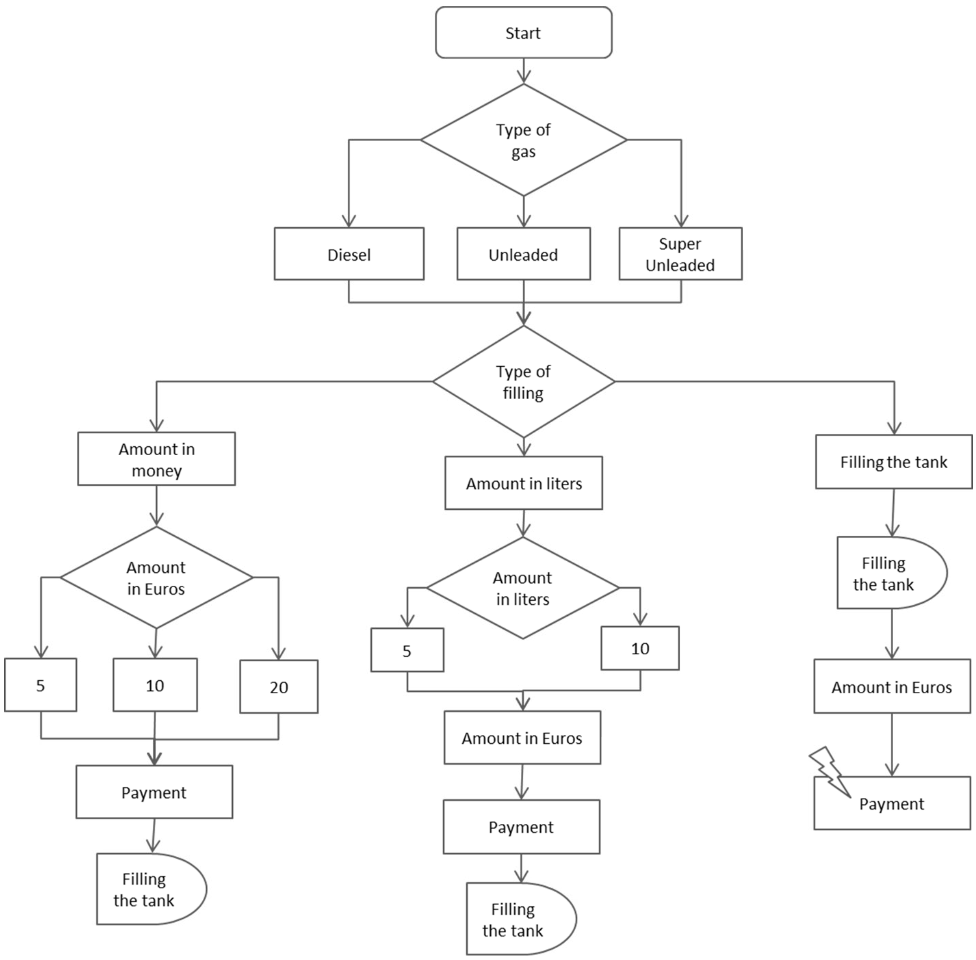


Figure 6.4. The flowchart of a typical fueling procedure catering for three alternative types of requests.

Payment in advance not only fails to satisfy the filling up requests, but also raises another issue. What if a driver requests 50 euro worth of fuel but the car's tank is actually topped up with less? Note that in a conventional fuel station this eventuality would raise no issue since payment always comes after fueling. You could of course incorporate a money refund process; however, such a process makes the system quite complex in terms of hardware involved, notwithstanding the need for constant monitoring and replenishing all kinds of banknotes. All the above possible situations are actually different user scenarios, as introduced in the previous chapter, which help us specify a more formal set of user requirements. User requirements act as a design space that let us imagine alternative ways to satisfy them.

Designing a new automatic fueling system requires making some head-on decisions, i.e., which/ how many scenarios are we going to cover. Should we, for instance, cover tank filling requests or avoid providing this option? On the one extreme we could imagine a very quick and easy fuel request dialogue, with one button for fuel type and one button for 5 euros worth of fuel clickable multiple times to reach the desired amount. No filling up option and no refund if the tank gets filled before reaching the selected amount. On the other extreme, we could suggest a sophisticated system with customized prepaid cards providing many options (e.g., selecting exact amount in liters, filling up the fuel tank etc.), but with the disadvantage that it only addresses to the needs of loyal customers. Both systems, among others, could be viable depending on the marketing strategy and the target audience of the company. Alternative solutions can be represented graphically, so that decisions on their architecture can be made before starting to design an interface for each of them.

You might notice that, like in the bathtub example, the upper levels of this analysis are identical to both the alternatives and the conventional system. This is perfectly normal, since the primary task itself (i.e., buy fuel) remains unaltered no matter the technological implementation. Different design solutions only alter the lower-level tasks of the process as we will see in the following section.

6.3 Hierarchical Task Analysis

Hierarchical Task Analysis (HTA) is one of the most popular techniques for representing how a task is performed, given the constraints imposed by an existing or future work system. HTA is based on the idea that each task can be broken down into task components at a number of levels of description, hierarchically structured and arranged, with more detail of the task being revealed at each level downwards.

In order to understand its function, we should begin by analyzing the actual steps we take when trying to accomplish a certain routine task. Take, for example, cash withdrawal from an ATM (Figure 6.5). First, we place our card in the slot, we type the PIN, we select the option "withdrawal", we type or select the requested amount, we remove the card and finally we collect the money from the slot. By reflecting on why each step is needed, we can formulate a higher level of task description being more abstract, and thus less dependent on the current lower-level implementation.

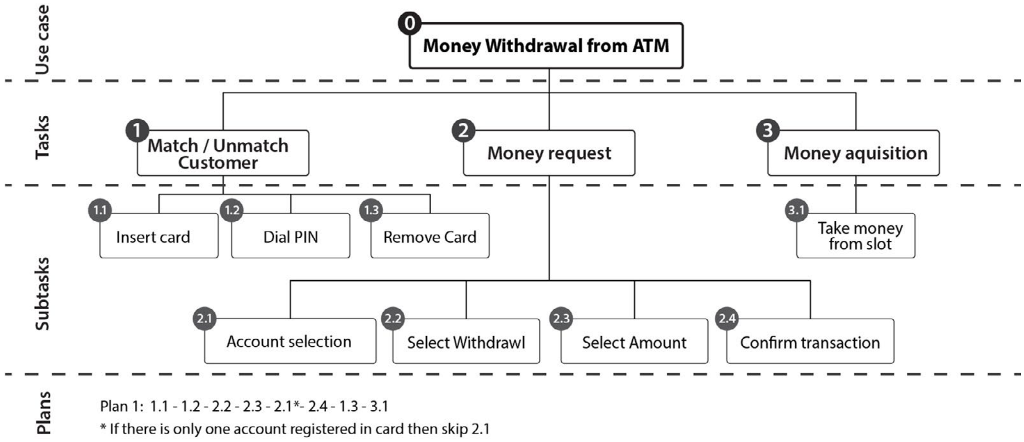


Figure 6.5. The Hierarchical Task Analysis for the task of withdrawing money from an ATM.

In our example, placing the card in the slot is needed for matching the card with a certain customer and their bank account/s. The PIN is requested for the identification/authentication of the user (PIN stands for Personal Identification Number). So, these two actions consist of a certain design implementation to achieve matching with some bank account(s) and then identification/authentication of the owner of the account(s). By understanding their role in the process, we can think of alternative ways to achieve the same demand. For example, instead of a PIN we could use biometric data, like fingerprints or retina scanning or sending a temporary code in the smartphone of the account owner.

All these design solutions have pros and cons; with the biometrics solution there would be no need to memorize numbers, but you could no longer ask your spouse or a friend to accomplish the withdrawal for you. However, the need for matching with a bank account and for verifying deliberate use by the owner remains intact so the upper-level tasks are independent of the technology used.

We should also question the entire section of matching with a customer for several services offered by an ATM, like the payment of a bill with cash or simply the browsing of the menu, to see all the provided services. In these cases, it seems unnecessary to pass through identification/ authentication steps, yet most contemporary ATM interfaces require so. Finally, the request to remove the card from the slot before opening the money tray (Plan 1: ... - 1.3 – 3.1) is deliberately designed as an interlock to prevent forgetting the card after receiving the money. In an alternative design solution for matching and identification, without the need of a card, this extra step (1.3) which causes waste of time for the user would be unnecessary.

Technically, an HTA is developed as shown in Figure 6.4. At the top of the HTA tree the main goal is placed (or in other words the general task). Then, by asking the question "how is this goal achieved?", the main subtasks that must be implemented to achieve it are recorded at the immediately lower level of abstraction. The development of the HTA tree continues for each subtask, using the same process as above, at the immediately lower level, until further analysis is considered non relevant. After all tasks and sub-tasks are in place, then use plans are added to dictate the order in which the goals are achieved. Use plans are numbered steps by which each task or sub-task will be carried out (Ashley French et al., 2019). If some tasks or sub-tasks must be completed in a specific order(s), due to work system constraints, then a plan to be followed is noted (using flowchart notations for cases where the plan includes conditional actions).

HTA is not prescriptive in the number of levels allowed, and utilizes heuristic "stopping rules" to help the analyst decide when analysis is complete (Annett & Duncan 1967, Stammers & Shepherd 1995). This depends on the goals of the work analysis and the judgment of the analyst. Also, another point on which the analyst is asked to decide, is how many and which possible situations will be included in the HTA. Thus, if the goal of the work analysis for withdrawing money is the design of the user-ATM interface or the usability evaluation of an existing interface, then the situations of lack of bank notes in the ATM, retention of the user card, inability to provide the requested amount due to exceeding the account balance, etc. should also be included.

So, through HTA, the designer can define the system in more abstract terms and therefore, decide on which implementation has fewer steps or better meets user needs. It must be noted that the HTA is quite different from a Flow Chart with which it is usually confounded, as the former structures tasks primarily in terms of taxonomical hierarchy while the later only in terms of sequencing. Below (Figure 6.7) you can see a Flow Chart for the same ATM money withdrawal process as the above HTA example.

Bear in mind that an HTA represents a certain process through the eyes of one type of user. If multiple types of users (or personas) are involved, then each one should have his dedicated HTA; multiple HTAs may, then, be combined in a workflow analysis (see below under Workflow Analysis). Nonetheless, secondary user roles, who act as facilitators to the end-user tasks, can be embedded in the end-users' HTA diagram, as parts of the system functions, and be analyzed separately in a dedicated HTA of their own.

6.4 Flowcharts

While the HTA describes the different levels of a task structure, a flowchart represents a workflow or a step-by-step process of executing a task (see Figure 6.7 for the ATM example). Therefore, unlike HTA diagrams, flowcharts do not analyze a task in levels of abstraction, but visualize detailed action paths or processes, inputs / outputs and decision points at the lowest level of task execution. In conceptual design, flowcharts are typically used after a basic concept of a future system has been decided upon (e.g., through HTA) to represent detailed action sequences in tasks requiring many steps with alternative paths. Flowcharts follow a well-known notation that makes use of box shapes denoting different types of actions or steps in a process, the most common of which are shown in Figure 6.6.

Figure 6.6. The most common box shapes used in flowcharts.

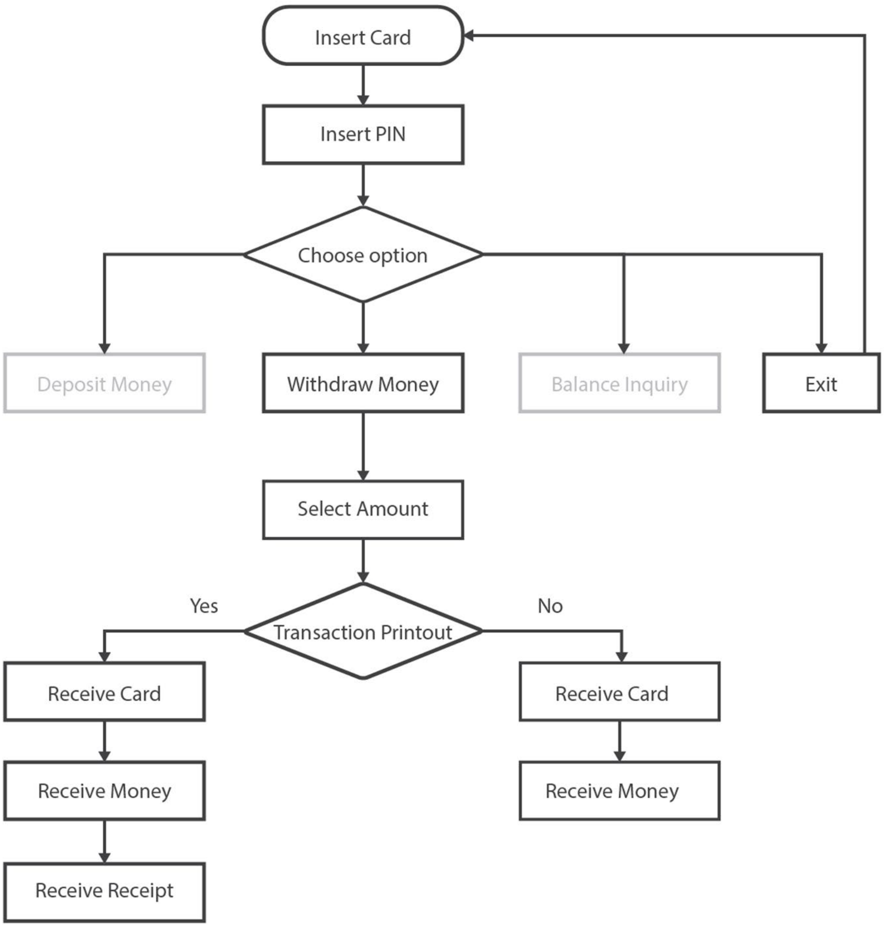


Figure 6.7. An example of the procedure of money withdrawal from an ATM so that you can see the differences with the previous diagrams.

6.5 State Transition Diagrams

Once user tasks and task flows have been specified, the next natural step is to consider the system interface aimed to support these tasks. Before dealing with dialogue and graphic interface details, it is wise to design/ specify interface behavior at an abstract level. A widely used method for this is through State Transition Diagrams (STD). An STD unravels all the different states of an interface (typically of a digital system), and all possible transitions among them. Transitions, also, designate the events or user actions under which the interface changes state during use. In short, through STD diagrams we can visualize the various states of a system interface and possible user, or system activated transitions. Such diagrams, besides being a powerful tool for the design of interface functionality and behavior, can also serve as a detailed specification of system behavior to be passed over to the software development team. Tentative STDs are also used at the early stages of the design process, where low fidelity prototypes must be made and tested.

A simple example of a State Transition Diagram is depicted in Figure 6.8, applied to the initial screens of a typical ATM interface.

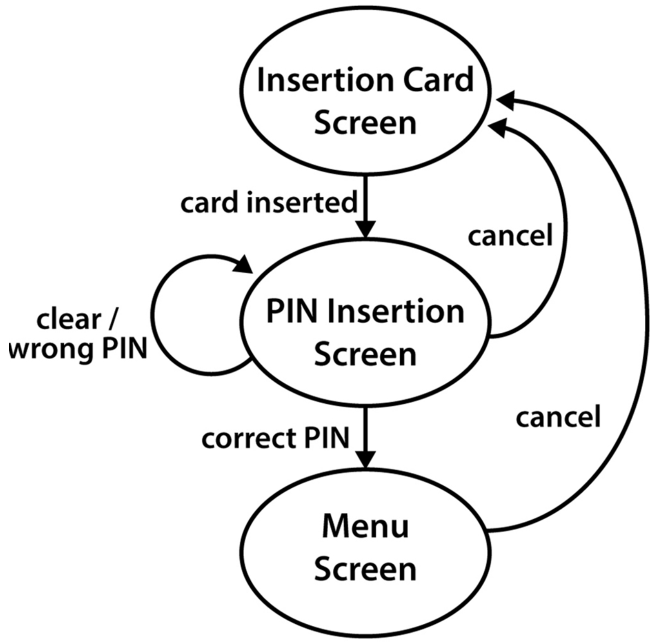


Figure 6.8. A State Transition Diagram for the initial screens of a typical ATM.

The STD in Figure 6.9 depicts the sequence of the three initial screens, but also specifies what happens when making a selection from the menu (user action) or even when entering a wrong PIN.

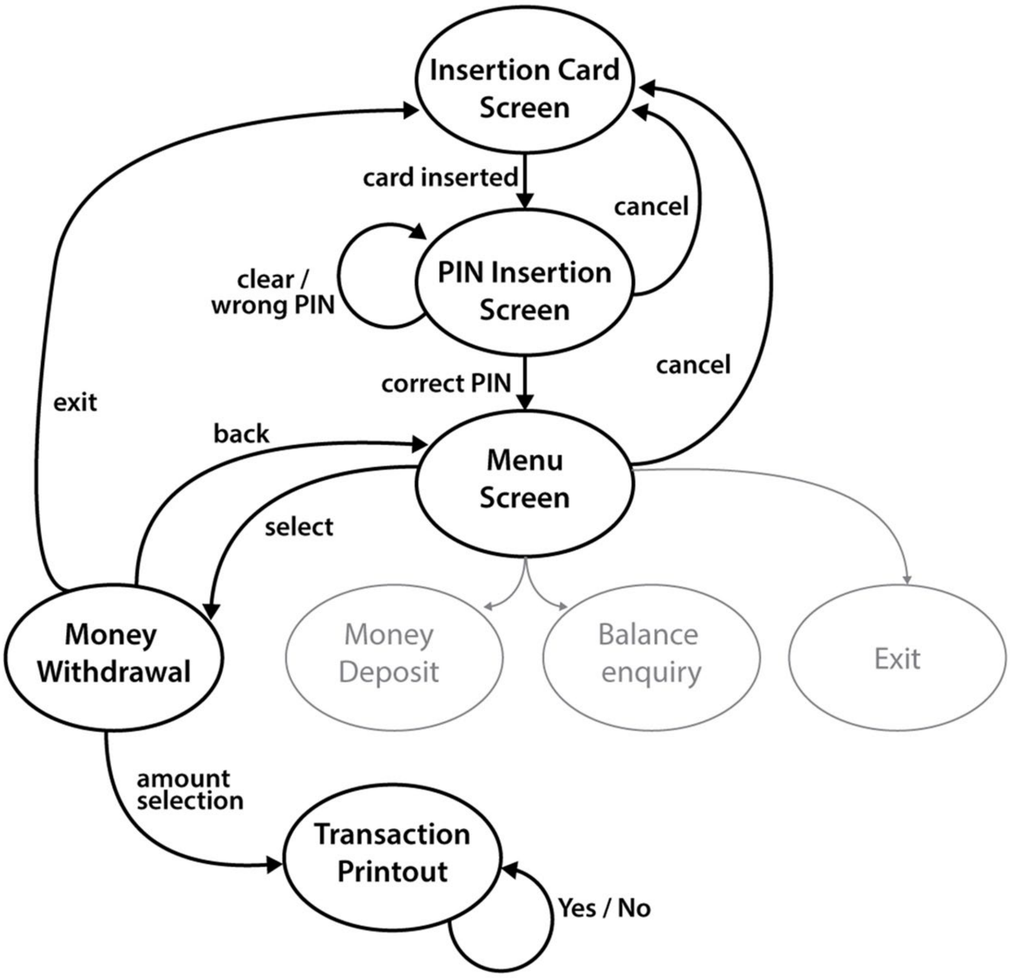


Figure 6.9. Partially developed STD of an ATM for the money withdrawal process.

Following from above, when reaching the main menu screen, the user can choose among several options available. For simplicity reasons, only the “Money Withdrawal” option is analyzed further.

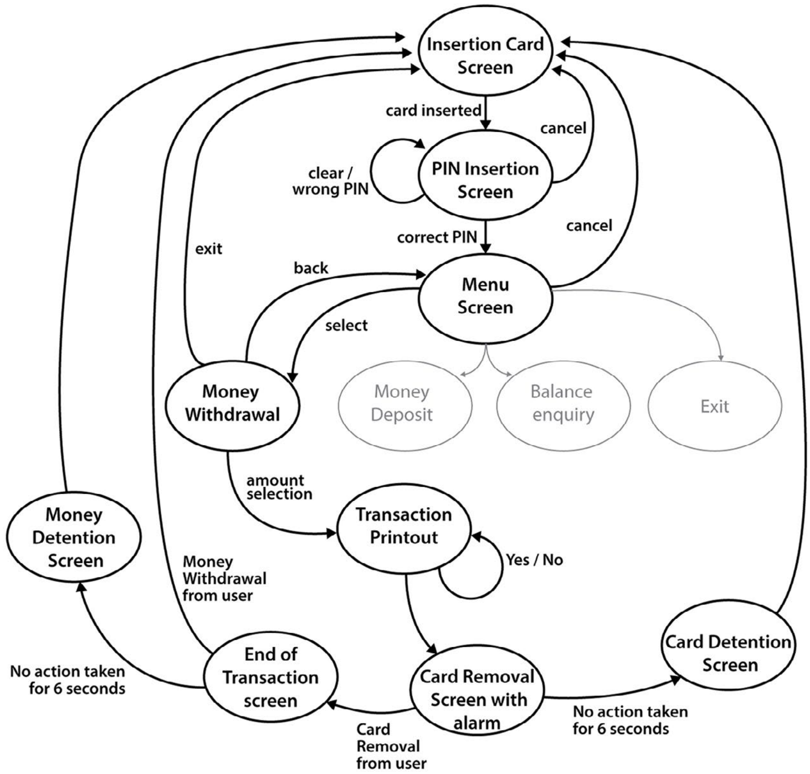


Figure 6.10. Fully developed STD of an ATM for the money withdrawal process.

Once selecting the desired amount of money, the system asks whether there is need for a transaction printout; then a series of checks are performed to ensure that neither the card nor the money will end up in the wrong hands (Figure 6.10).

Overall, STDs serve as an overview of the states and reactions of a system, making it easy for the designer to envision various shortcuts and simplify certain processes. As such, they directly inform the development of low fidelity prototypes for early user testing, discussed in the next chapters. Finally, as already noted, fully developed and validated STDs are also great means for communicating system dialogue architecture to the system development and graphical user interface implementation teams.

6.6 Workflow analysis

When there is a need to visualize the internal processes of an organization or a service provider, we make use of a Workflow Analysis Diagram. This tool helps us improve the operational efficiency of specific business processes, by identifying redundant actions, bottlenecks and inefficient usage of resources. In these diagrams, all people and teams involved are included along with the artifacts and information being circulated among them. For instance, in a Workflow Analysis of the outpatient care process, in a healthcare system, we would include appointment scheduling, patient history, medical actions, clinical results, the medical staff involved as well as the administrative checks. Through Workflow analyses we can identify system level inefficiencies, such as potential delays in a medical procedure as well as possible points where errors can occur (e.g., at the transcription points of medical data). Being mostly oriented in detecting inefficiencies in corporate environments, we will not go in a deeper analysis of this tool, but it is important to acknowledge the circumstances where it is mostly suited. A simplified workflow diagram, however, is presented in Figure 6.11, concerning the process of an examination during pregnancy. The information flow and the internal processes across organizational units of a medical center are depicted, so as to unveil potential errors or delays.

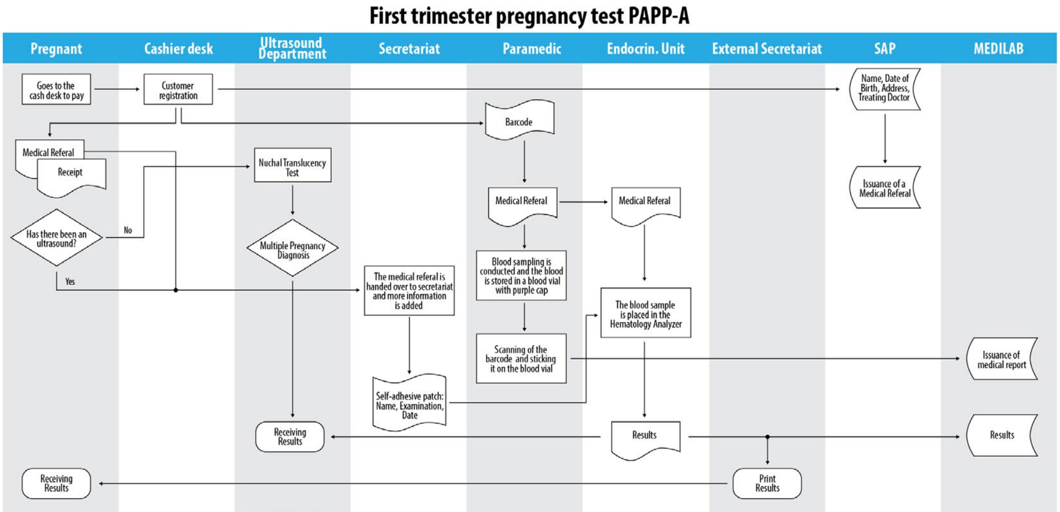


Figure 6.11. A workflow analysis of an examination during pregnancy in a medical center.

6.7 User Journey Mapping

A User Journey Map (UJM) is a visualized storytelling technique which helps us follow the "journey" of a user trying to achieve a goal through a service provider. Typically, a service contains multiple means and points of interaction with a user (e.g., physical visits, phone calls, digital interactions). Used mostly in Service Design, the Journey Mapping creates a holistic view of user experience, by bringing together and visualizing all the "touch points" (i.e., interaction points) of the user with the system, through a scenario, along with the potential emotions that occur during this journey. The main purpose is to understand the user's needs and address them accordingly.

Journey maps can either be current-state or future-state, depending on whether we want to visualize the user experience of an existing system or the ideal user experience of a future system. The former serves as a tool to identify existing problems, while the latter to help us create new experiences. It is generally advised to always start with a current state Journey Map to have a robust evidence-based scenario, and then if needed proceed with a future-state where more assumptions will be made.

In practice, UJMs utilize a "persona" (i.e., a fictional user) who will follow a certain scenario to achieve a goal through our service. In this imaginary journey, we must record all the interactions of that persona with the service, as well as the emotions that each interaction might trigger. In each of these phases we will have to record the insights and discover the opportunities for improvement by asking questions such as: "Were all the steps necessary?" "Could there be done in a more enjoyable way?" "Does our service respond as expected from the user?"

An example of a simple UJM concerning a public service, in this case the registration of a passenger car and acquisition of new plates from the Greek Transport Office is shown in Figure 6.12.

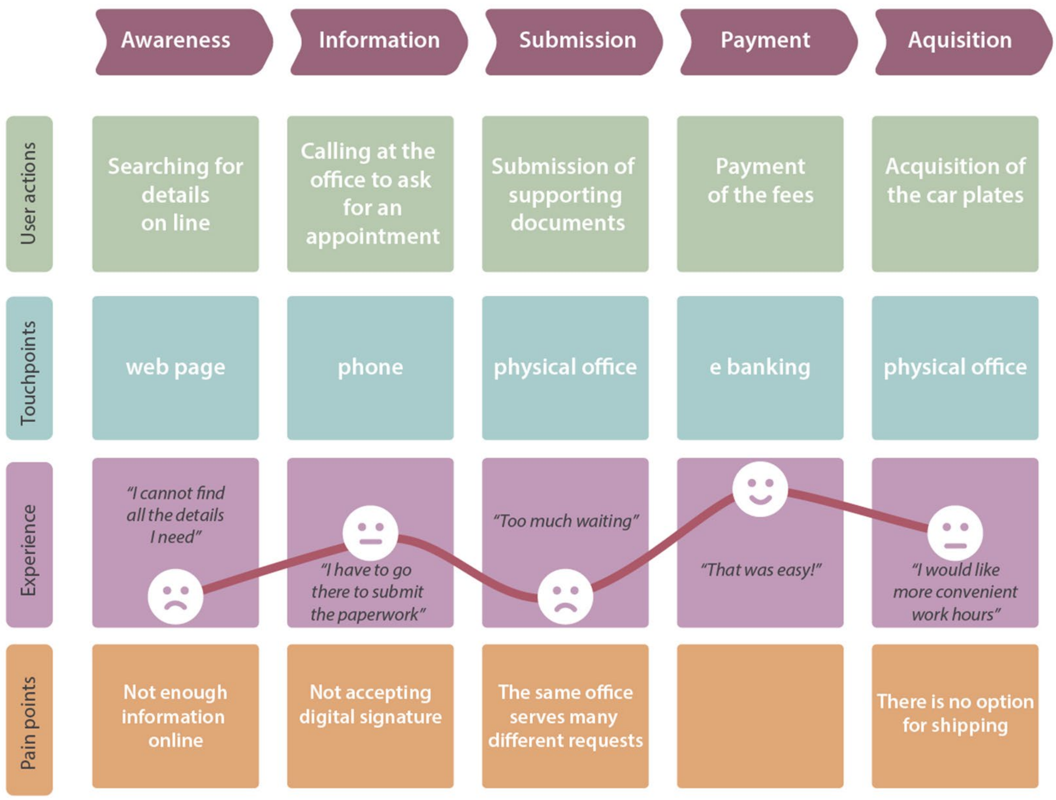


Figure 6.12. User journey map for the registration of a passenger car and acquisition of new plates.

6.8 Service Blueprints

In Service Design a blueprint visualizes the organizational processes that take place in order to provide a certain service to a user or a customer. The goal is to better organize the resources of a service provider to improve the customer experience, as well as the employees' experience. Contrary to a User Journey Map, a Service Blueprint (SB) includes all the people involved, the physical and digital evidence, along with the front-end and back-end processes. As Sarah Gibbons (2017) from Nielsen / Norman group explains "Blueprinting is an ideal approach to experiences that are omnichannel, involve multiple touchpoints, or require a cross functional effort (i.e., coordination of multiple departments".

Usually, an SB follows the development a UJM, and incorporates a simplified of the latter in the form of the user's actions as its top layer. Next, there is the front-end action layer which describes the direct interactions of the company with the user, whether physical, digital, automatic or human initiated. The third layer consists of all the backstage actions which take place beyond the user's visibility or knowledge, and concern organizational protocols. The last essential layer of an SB contains the support processes that happen behind the line of visibility of all the above participants. Apart from the recognition of these layers, it is also important to connect the above elements accordingly, to show their dependencies and the route of the data after each triggering action. Lastly, some extra elements could be digital or physical evidence which circulate among the participants, some timestamps giving the duration of these processes, and the annotations of some of the elements depending on our goal (e.g., pleasant, annoying etc.).

Figure 6.13 presents an example of a blueprint for the same service as in the above Journey Map (Figure 6.12). You will notice the different lines of visibility to the end-user which enables the designer to dig deeper in the organizational processes of the service provider. These diagrams can be as analytical and detailed as needed, and with as many extra elements that would help us spot the bottlenecks or the shortcomings of the system.

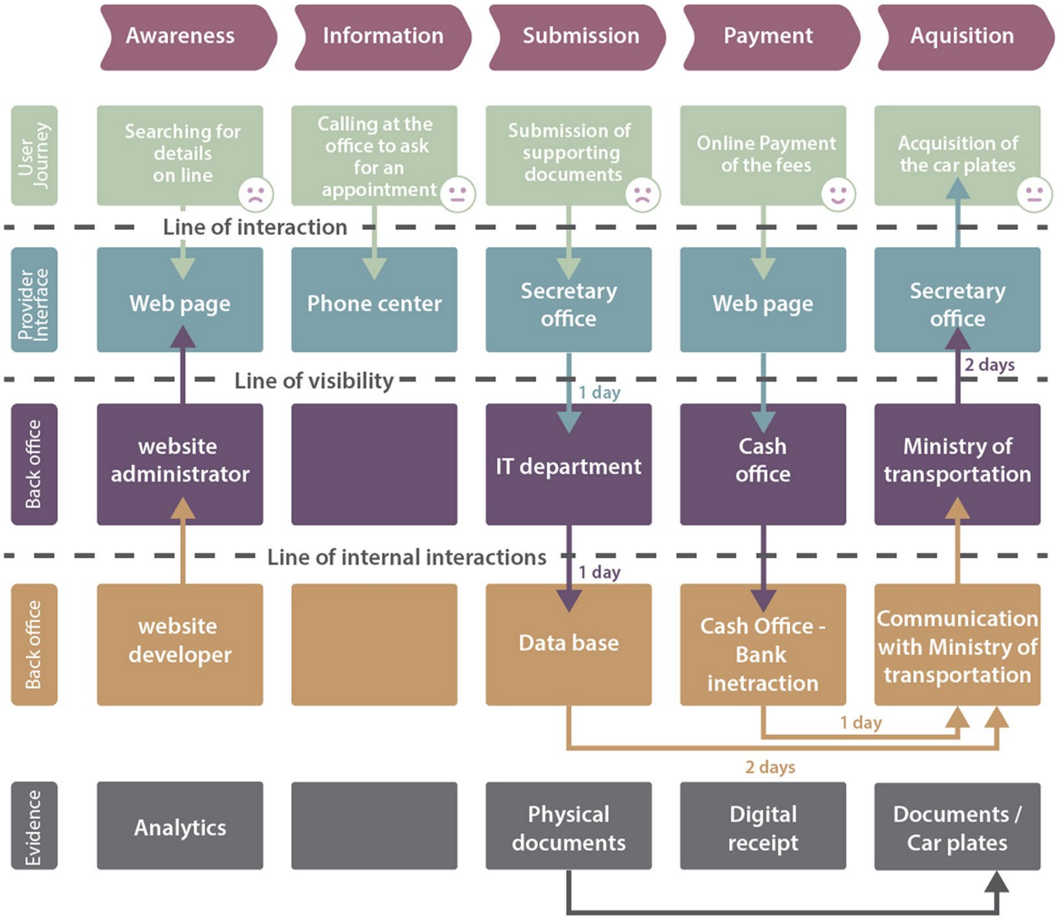


Figure 6.13. A Service Blueprint following the User journey map for the registration of a passenger car.

6.9 On Ideation and Brainstorming

Ideation, i.e., the generation of ideas, is a quite vague process that entails affective, aesthetic as well as social skills that require specific training best obtained in design schools. Nevertheless, various techniques exist to facilitate ideation, namely brain storming, mind mapping, challenging of own assumptions up to the use of random words as a probe imagination of novel concepts. Most of these techniques aim at enhancing team inventiveness by providing some structure and discipline to an otherwise messy process. These will not be presented in detail aside from a brief note on some ideation fundamentals and common pitfalls. For more comprehensive information on ideation methods the interested reader should refer to specialized textbooks.

Above and beyond specific techniques, there are some key elements that have been proved to help this fuzzy process, and are mostly targeted in overcoming some common cognitive and affective biases.

One of these key elements is the value of polyphony during the ideation process. The work in groups consisting of heterogeneous individuals in terms of age, gender, personality and cultural background has been proved far more creative and providing more innovative solutions than the individual ideation process, of even the most experienced designers. In recent years, the term participatory design has gained ground in corporate environments, where such ideation groups also include end-users or other stakeholders whose different points of views is an invaluable source of information. Of course, in such ideation groups the designers do not seek final detailed solutions, nor are they obliged to follow the ideas proposed or to value all opinions equally. The main purpose is to help expand their dimensional horizon of the system. The responsibility for the final design always lies exclusively on the designer(s), no matter what the group has proposed.

Since the main goal in this participatory process is to broaden the ideation base, it is essential for each participant to feel free to express ideas without being criticized or mocked. Refraining from criticism on others expressed thoughts or ideas is the first and most important rule in the group ideation family of methods called "brainstorming". In brainstorming, each idea, no matter how extravagant or irrelevant might sound, must be freely expressed, so as to trigger the group creativity. It is the moderator's responsibility to delimit the scope, or if necessary, to prevent negative comments and try to extract the useful side of each idea expressed. A way to balance this tendency of directly criticizing others' ideas is to write them down anonymously, and then simply vote on each of them as a way to promote the most popular ones. After this distillation process the preferred ideas are redistributed among participants for further brainstorming.

Another common phenomenon observed mainly in novice participants is their difficulty to overcome fixation on their own initial idea. Since the brainstorming process does not necessarily produce direct solutions, the participants' ideas must be brief and rough. Not all technical details must be solved at that point, neither should these solutions be very detailed. However, it is inevitable that some participants will fixate too much on a solution that they came up with and try to capitalize on this idea with more and more details. To avoid this psychologic bias on focusing too much on our own ideas, some brainstorming techniques have been proposed, such as the Group Passing technique. According to this, the participants work either individually or in pairs to produce ideas which are then transferred to the next participant or subgroup who are asked to develop on these ideas and so on (Figure 6.14). Note that handing over one's primary ideas to other participants for further development inhibits the natural human tendency of criticizing and enhances the feeling of group ownership. In this way all the final concepts lack individual ownership as they are technically products of teamwork.

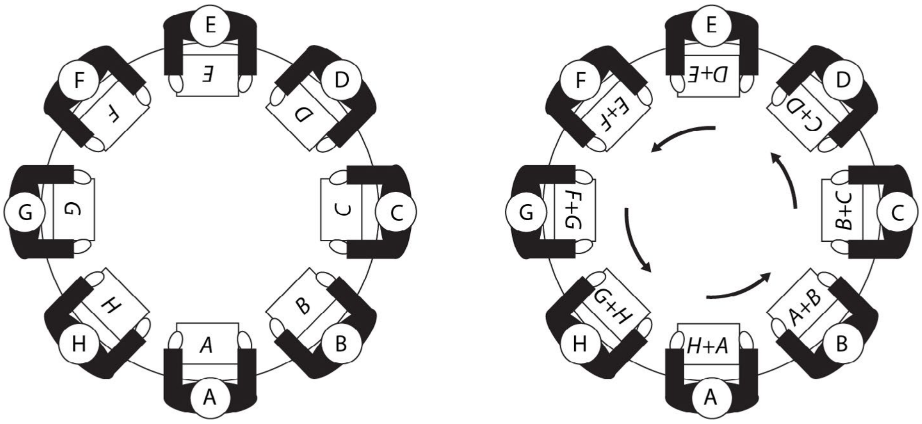


Figure 6.14. Group Passing Brainstorming.

Even the most common brainstorming technique (called Team Idea Mapping method) focuses on promoting the collaborative work over the individual ideas. In this method all individual ideas are written on sticky notes anonymously, and then placed on a big board to form an idea map. During this consolidation phase the participants develop a common understanding of the problem space and new ideas are produced through association of the initial suggestions.

References

Annett, J. & Duncan, K.D. (1967). Task analysis and training design. *Occupational Psychology*, 41, pp. 211-221.

French, A., Taylor, L. K., & Lemke, M. R. (2019). Task analysis. In *Applied human factors in medical device design* (pp. 63-81). Academic Press.

Gibbons, S., (2017, August) *Service Blueprints: Definition,* Nielsen Norman Group https://www.nngroup.com/articles/service-blueprints-definition/

Stammers, R. & Shepherd, A. (1995). *Task Analysis. In Evaluation of Human Work*, J. Wilson & N. Corlett (eds.). London: Taylor & Francis Ltd, 2nd edition, pp. 144-168.

Chapter 07:
Prototyping / Iterative Design

Chapter 07: Prototyping / Iterative Design

Chapter Summary

The chapter presents iterative design, i.e., multiple prototyping with escalating levels of specification and subsequent user testing as the preferred detailed design approach. Specifically starting at an early stage, paper prototyping and physical mock-ups are discussed followed by subsequent prototyping methods such as Wizard of Oz and Wireframes up to semi functional systems. The above methods / tools are demonstrated through proprietary and literature examples alongside suggested classroom exercises that the authors have tested with students of various disciplines. Furthermore, the most widespread usability inspection methods are presented namely Heuristic Evaluation and Cognitive Walkthroughs. Next, an introduction to experimentation with users is presented along with a broad reference to different user testing methods, i.e., the various alternatives of Usability Testing and an introduction to the Eye-tracking method for a more in-depth analysis of the user's action. Finally, post launch monitoring is discussed as a means of evaluating efficiency after system deployment and for scrutinizing eventual retrospective design interventions along with A/B testing as a hypothesis testing method for existing designs. The above methods are evaluated on the basis of their usefulness in specific requests concerning the maturity of the designed system and the available resources.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of previous chapters.

7.1 Striving for usability

The practical aim of Human Machine Interaction is to enhance / optimize the coupling between human users and their machines. The quality of such coupling is evaluated through the concept of "Usability". According to ISO 9241-11:2018, "Usability" is defined as the extent to which a system, product or service can be used by specified users to achieve specified goals with *effectiveness, efficiency* and *satisfaction* in a specified context of use, where:

- effectiveness is the accuracy and completeness with which users achieve specified goals
- efficiency is the resources used in relation to the results achieved. (Typical resources include time, human effort, costs and materials)
- satisfaction is the extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations. Satisfaction includes the extent to which the user experience that results from actual use meets the user's needs and expectations. Also, the anticipated use can influence satisfaction with actual use.

As all the above aspects are inextricably linked to specific users, and specific goals, it is rarely possible to achieve high levels of usability based only on theoretically derived design principles. Although such principles are indeed important and have been tackled in previous chapters, in any design endeavor, empirical testing with prospective users is essential. This calls for a process of progressively shaping a design through successive design cycles iterating among ideating, prototyping and testing stages. In this chapter we will introduce several methods aiming to guide the designer through these successive design cycles depending on the given task, the specified population and the available resources.

7.2 Prototyping

"*Fail early in order to succeed sooner*" is a quote in the field of innovation design, attributed to David Kelley, founder of the design and consulting firm IDEO. Not only sooner we would reckon but also at a lower cost. The cost of changes on a new product, whether physical or digital, increases significantly with each successive phase of development. As we have seen until now, the first phases of product or system development consist of investigating, analyzing, and imagining/planning. These initial phases require relatively small resources compared to the latter stages of detailed design, or software coding. When detailed design starts, the resources needed to build and code a system increase exponentially so that any needed change at these stages will exponentially increase time and cost (Figure 7.1).

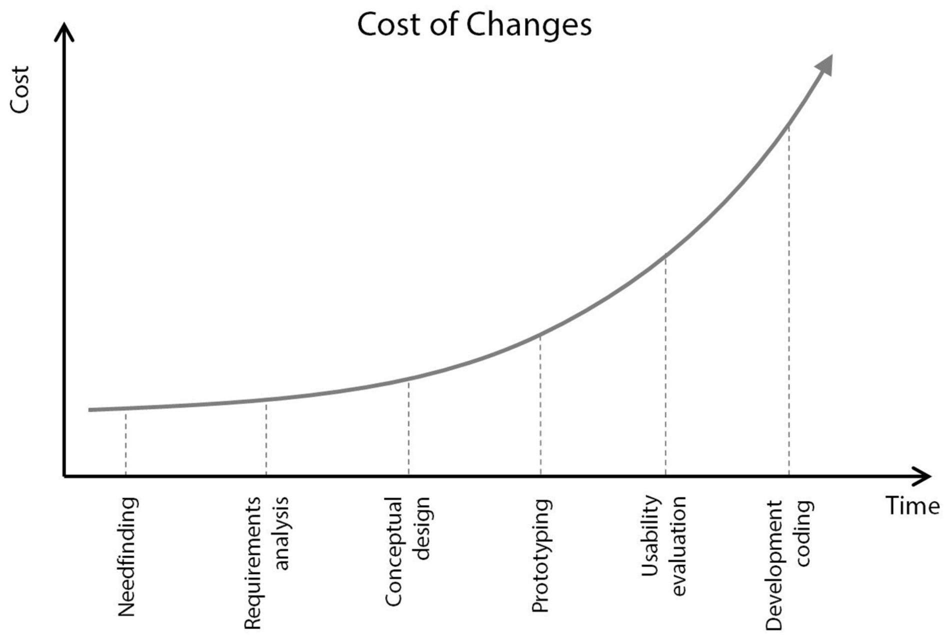


Figure 7.1. Cost of changes in new product/ system development increase exponentially as the development process unfolds.

To avoid increased costs and time delays, several stages of draft designs and small-scale testing are highly recommended. Such small-scale design and test cycles allow us to check the effectiveness of our ideas, and to reject or correct them with relevant ease. What's more, by allocating resources on creating several rough designs instead of a single high fidelity one, we can test various alternatives that otherwise would have been impossible to compare. In what follows, we present several design/ testing tools that have proved useful in the early stages of the design process.

7.2.1 Storyboards

Storyboards (much like in the movie industry) consist of sequences of sketches that try to convey the basic elements of how the intended product or system is envisaged to be used in practice. Storyboards are frequently employed at the start of innovative interactive systems to help the team members imagine and converge on the basic concept. In addition, the final storyboards of a design can also be used as promotional material to various stakeholders or potential clients as a way to acquire funding. The main idea is to visually communicate a small story illustrating the problem that the design team will try to address and the main idea of the potential solution, through a series of chronologically arranged frames. This is usually done through the eyes of a character (persona) in a way that creates empathy and makes the situation easier to grasp.

Technically, the sketching of the frames should be above all clear, while the artistic quality is irrelevant when they serve internal purposes. In most cases, captions that either externalize the character's thoughts or describe the situation are added in the frames. A good practice is to keep the story short, to communicate the central idea in a fast manner. It is also very common to use a small introductory explanatory text when handed over among members of the team or external interested parties. Sometimes storyboards are used in service design to enrich journey maps (see Chapter 6) by visualizing each interaction of the user with the service in question. Figure 7.2 presents an example of a storyboard.

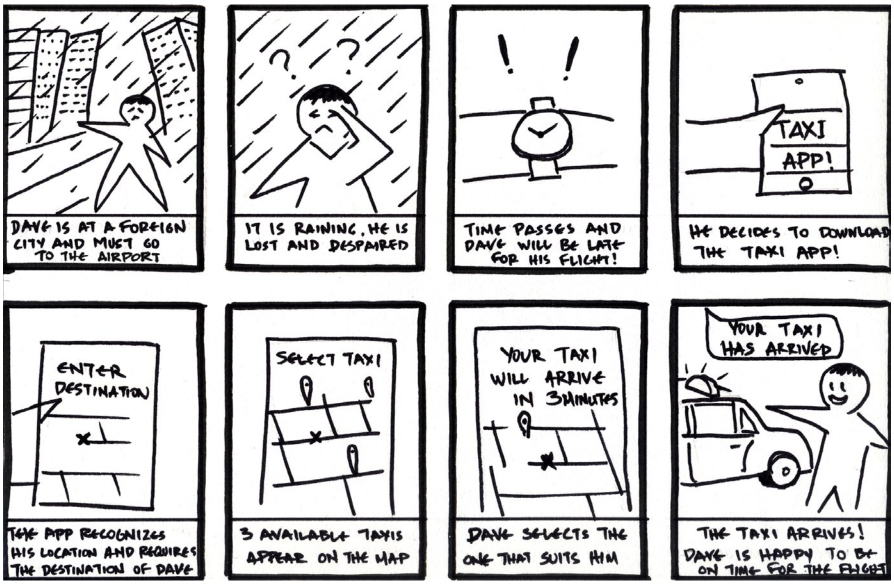


Figure 7.2. Storyboard depicting a taxi service application concept.

7.2.2 Paper prototypes

In interactive system development, it is generally advised to start with nonfunctional prototypes, and progressively add functionality, conducting iterative user-tests along the way. The most straightforward type of early prototyping is the “paper prototype”. Paper prototypes are representations of the interface of a digital product, sketched on paper along with a basic interaction structure. Besides being a versatile method for idea generation, they are typically employed for early user testing to validate basic navigation structure and wording.

Paper prototypes should intentionally look rough in terms of drafting quality and materials used (Figure 7.3). On the contrary, they should be readily disposable and redesigned even during the testing process. This is important for at least three reasons: (i) to save time and resources, (ii) to prevent designers from getting too attached to a prototype and (iii) to better communicate their provisional purpose to the test participants. Indeed, sketch roughness and cheap materials accentuates the provisional character of the design helping them to focus only on the essential elements at this phase, suggesting also that potential changes are easy to make –even sought for. This helps avoid a typical tendency of test participants to please the designers when being tested (a tendency we often see when using relatives or people of our close social circle as testers). Note however that although the prototypes should look rough, their physical scale should match the scale of the final interface to correctly assign the appropriate amount of information.

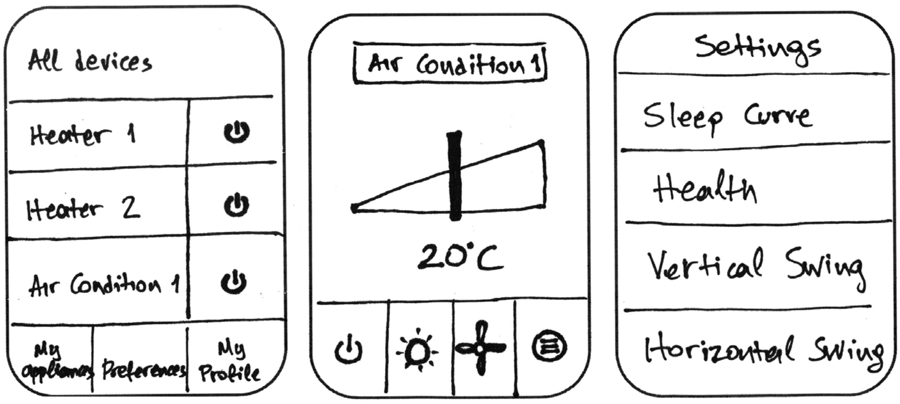


Figure 7.3. Paper prototypes of a smart home application. The design is rough and comprising only the available choices and their arangement on the screen.

7.2.3 Wizard of Oz

When conducting user testing with paper prototypes one faces the problem of how to animate a dummy interface with no embedded operating system. Here the "The Wizard of Oz" technique comes handy. In the well-known novel and movie bearing the same name, the mighty wizard character was nothing more than a mechanism with switches and levers controlled by an ordinary man, giving the impression of a living machine. The same idea of replicating the automation is used to simulate the behavior of an operating system. Paper prototypes are usually animated using this technique so as to give a similar experience one would have when using the real interface.

In practical terms, a moderator gives a certain scenario to the participant, and then takes the role of the computer algorithm that interconnects the different screens and presents the respective elements depending on the user's actions (Figure 7.4). For example, when the participant makes a selection on the home screen, the moderator, holding all relevant cards, presents the corresponding card. The same happens when the user selects a dropdown menu that must open on top of the previous card. The animating task is quite demanding, especially when there are numerous active elements; the moderator must be familiar with all the intended functionality and act fast to minimize idle time after each selection. It is important to note that the moderator, apart from presenting the respective elements, must not intervene in any way in the testing process (e.g., by helping or preventing the user's actions). All sessions must be conducted under the same scenarios prepared in advance according to the guidelines given in a Chapter 5 (e.g., include two or more specifying elements, describe situations not very obvious). It is essential to record all sessions with a camera and microphone for subsequent analysis, e.g., in terms of time taken, committed errors, users' comments or any other metric that seems fit.

Figure 7.4. Wizzard of Oz: when the user makes a selection, the moderator must be ready to respond by revealing the next screen.

The limited technical resources needed to implement this method come with a cost in human labor. For each user test conducted, one moderator is needed to perform the animation task of presenting the cards, thus pragmatically limiting the testing sessions to only a few (4-8).

Despite its inconveniences, the Wizard of Oz technique is quite versatile and can also be employed in early prototypes of auditory interfaces. In this case a set of auditory messages are prepared along with the interaction structure of the auditory interface, and the moderator selects the appropriate message each time a test participant makes a choice, either enabled through speech or number selection on a keypad.

7.2.4 Wireframes

A more elaborate prototyping technique for visual interfaces is wireframing. Similar to an architectural plan, a wireframe is a structural outline of a system's interface, aiming to allocate the appropriate real estate to the composing elements along with their functionality or intended behavior. As in the paper prototyping, wireframes lack any styling, colors, graphics or even use of special fonts (Figure 7.5). Instead, the illustrations should focus on the arrangement of different elements on the screen, the wording of menus and the relationships within the active content. If there are text elements where the content is irrelevant to our validating purposes or not yet created, dummy text (e.g., lorem ipsum) is used just to fill the blank space.

Wireframes are typically created with digital medium, but they must look simple, two-dimensional and unrefined. Some wireframing design software even use "jagged lines" to make the design look rough and premature. The unrefined look of wireframes, besides the reasons discussed in the paper prototyping section, helps to separate interface functionality from aesthetic choices. Aesthetic interface design is usually implemented at the final stage of the design process based on requirements stemming from branding guidelines.

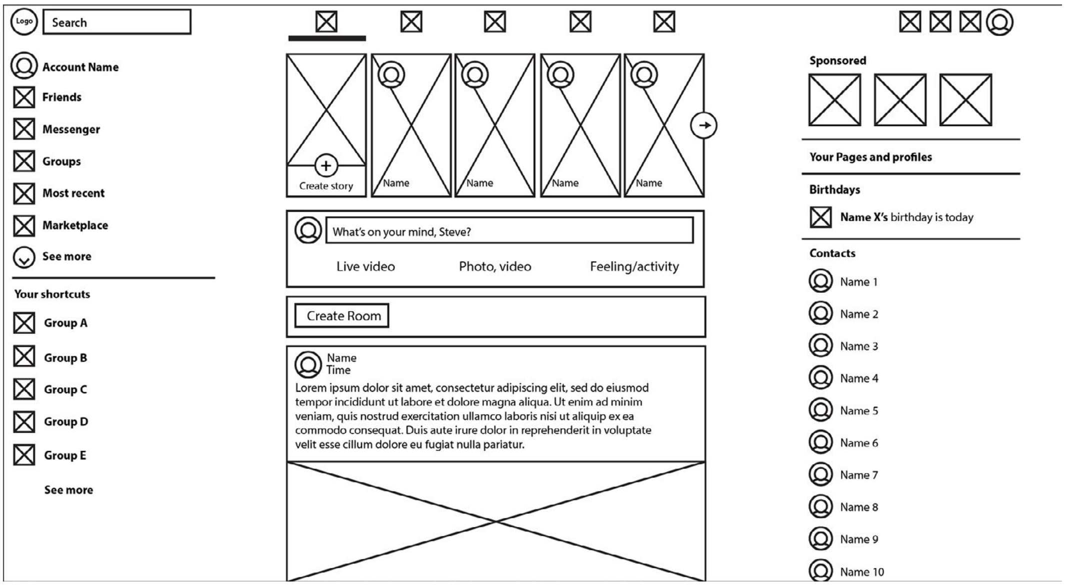


Figure 7.5. A wireframe prototype for a social-media homepage. Avoiding aesthetic choices, a wireframe concentrates on labeling, spatial arrangement and grouping of different elements.

Unlike paper prototypes, wireframes employ active links between menus screens, and content. The interaction functionality of these prototypes provides a more naturalistic testing experience to the user. Nevertheless, it inevitably needs more development resources, so it is generally employed at a later design stage, when initial issues have already been resolved. On the positive side, wireframe user tests can also run remotely, providing convenience for both participants and designers.

7.2.5 High Fidelity Prototypes

When most decisions on the interaction and interface structure have been made, the focus turns on the aesthetics of the design, i.e., its look, feel and interactivity. The dummy pictures and other user interface elements will be replaced with the final ones (buttons look and feel, fonts, colors, etc.). At this phase, all functionality, apart from the connection with back-end systems (e.g., databases), must be implemented so that the high-fidelity prototype can be tested with users on more complex scenarios (Figure 7.6). With such prototypes, more sophisticated testing methods like eye-tracking can be applied, since the added graphical elements will pervasively affect user perception. High Fidelity prototypes are a significant milestone in the design process as they bring us only steps away from a finalized product. Even though, design iterations might, still, be needed. In fact, the significant resources required for these prototypes means that interaction structure, navigation and functionality of the interface should be already finalized before starting development, as the cost for changes will be significantly higher at this final design phase.

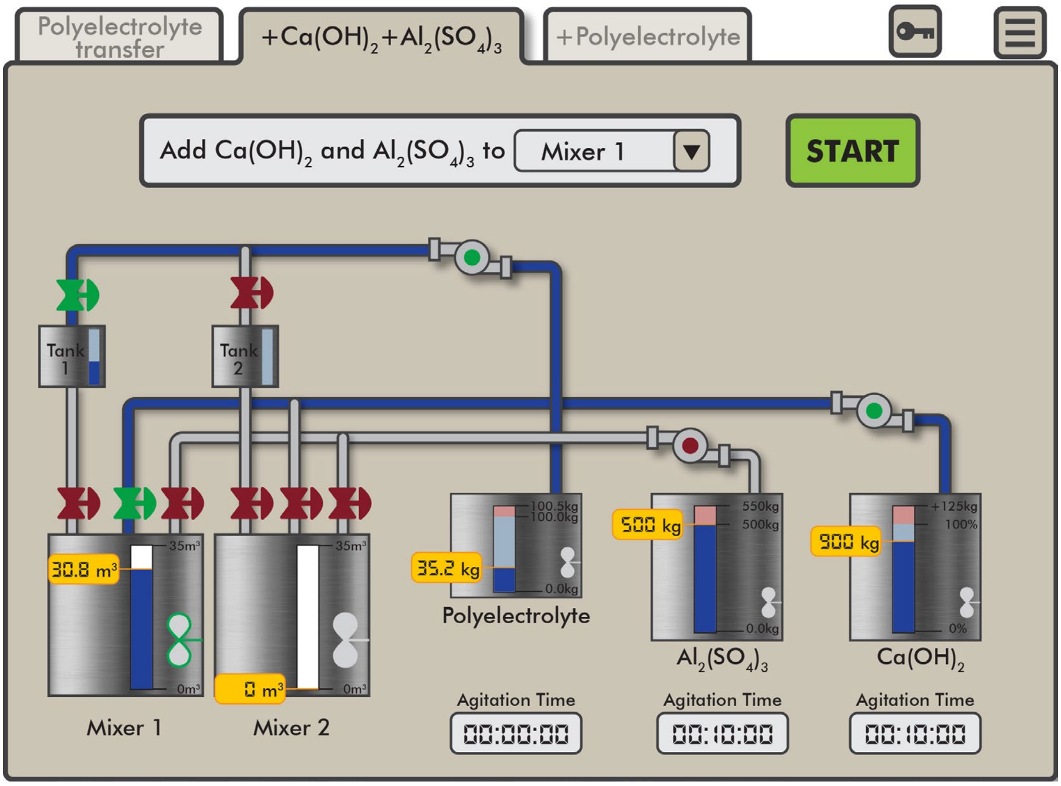


Figure 7.6. High fidelity prototype of a SCADA interface (mimic diagram) controlling a chemical process. The interface is fully functional, and the elements are designed in detail. Such prototypes can be put to test with actual prospective users under realistic conditions.

7.3 Expert usability evaluation methods

In User Centered Design, iterative evaluation and testing are essential throughout the design process. The least resource consuming, and easier to implement, is a family of methods called Expert Evaluation. Expert evaluation methods rely on “expert knowledge” to assess the usability of an interactive system. The term “expert” is used to designate individuals that have theoretical knowledge and training in interactive systems design. Expert evaluation methods, being inexpensive and easy to implement, allow fast assessment of alternative concepts after quick redesign iterations, mitigating rework costs during development. They are thus particularly suitable for the early stages of a design process. In small scale projects the role of expert evaluators can be taken by members of the design team themselves; however, in large scale and/or critical projects evaluators should be independent from the design team. Three well-established expert evaluation methods are presented below.

7.3.1 Nominal task evaluation

The simplest way to evaluate an interface is by simple quantitative task execution metrics, based on the nominal task execution. Typical metrics of this sort are the number of steps, or the nominal time needed to accomplish a certain task. These metrics can be carried out by the members of the design team, without the need of external participants, and are used to compare the performance of a trained user among alternative interface designs. For example, the number of steps necessary to accomplish a task, whether these steps consist of clicks on a screen or pushing of levers and buttons or even typing of plain text, provides an indicator of which design is more efficient when used faultlessly by a trained user. Despite being easily measured, this metric is not to be confused with actual navigation efficiency, as it does not take into account the possible errors which can be caused by the misinterpretation of labels or of the affordances and constraints of the system. For example, a certain task can be nominally accomplished with only three steps, but the wording/symbol of the first step might be so confusing that most users will navigate away, ending with considerably more steps before actually getting to it, while an alternative interface might require six steps, but it is implemented in a more error-proof way. Therefore, task-based metrics are best suited for evaluating interfaces destined to expert/profesional use.

A brief example illustrates this well. The main page of the Athens International Airport (Figure 7.7) features a widget trough which a user can directly reserve a car parking spot in just nine clicks. In a quantitative metrics evaluation, this parking reservation process would get a very good score compared to alternative designs from other airports. However, when tested with ten real users it was observed that none actually used the widget, but instead went straight to the menu on the right (green rectangle) and spent considerble time and actions, trying to find the parking reservation page among the four different choices in the drop-down menu: e-parking / airport parking / parking procedures / parking services. As it commes out, the nominal number of steps metric, although fast and inexpensive, is limited by its blindness to the final user point of view. Nevertheless, it is still valuable particularly for evaluating interfaces destined for expert/profesional use.

Figure 7.7. The main page of the Athens International Airport, featuring a parking reservation widget (red rectangle on the left) and a parking icon on the right (green rectangle).

The same can be said for the nominal task duration metric, which is typically used when task steps cannot be clearly distinguished. The real time taken to accomplish a task in user testing sessions may prove quite longer than the nominal as estimated by the design team. Nevertheless, nominal task duration is still valuable since it provides a solid baseline to overall interaction efficiency and will be one of the major quantitative metrics in most testing and evaluation methods throughout the development of an interface. Note also that nominal task duration is tricky to use for early prototype evaluation, as system response lags cannot be accounted for at this stage.

7.3.2 Heuristic Evaluation

Heuristic Evaluation (Nielsen 1993) is a well-established Usability Inspection method aiming at identifying generic usability problems in user interface design. It does not rely on users but to a small group of experienced evaluators, who review the user interface and judge its compliance to ten predefined usability principles. The evaluators, in addition to a good grasp of these principles, should preferably have academic training in Human Factors and Ergonomics or Human Computer Interaction.

The process of the Heuristic Evaluation includes the following steps:

1. A set of representative use-case scenarios of the system are developed.
2. For each use-case scenario, the evaluators inspect the interface independently of each other and record the usability problems they find. Alternatively, evaluators inspect the interface as a team and a team moderator records the findings. It is important to record all problems identified, regardless of whether or not there is agreement between evaluators.
3. The moderator processes the recorded problems and removes similar records.
4. The evaluators go through the list of problems independently of each other and rate or rank the problems according to their importance.
5. The moderator combines the scores of each evaluator, finding the average, and ranks the problems in order of importance (Figure 7.8).
6. The design team looks for solutions to the identified usability problems and estimate their implementation cost.

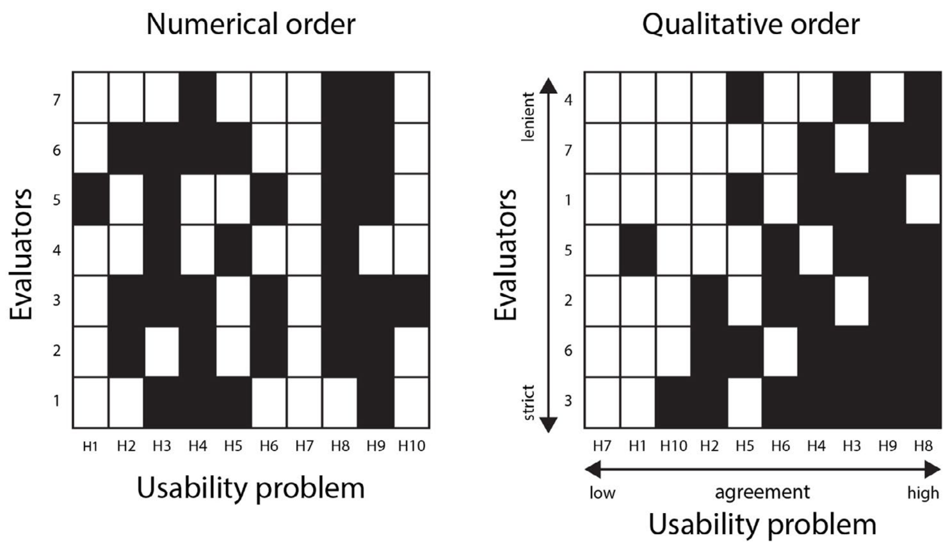


Figure 7.8. Not all evaluators are equally effective in identifying problems. The matrix on the left depicts the results of 7 evaluators in identifying usability problems according to the ten heuristics (H1-H10). Black squares signify that a particular evaluator found a particular usability problem. On the matrix to the right, the rows are re-sorted based on evaluator's strictness and the columns re-sorted based on agreement among evaluators.

Empirical evidence (Nielsen & Molich, 1990) has shown that not all evaluators are equally effective at identifying usability problems, nor do they identify the same problems (see Figure 7.8). Consequently, the question arises as to how many are necessary to be employed each time. Based on cost/benefit analysis, the best results have been found to be obtained from 3 to 5 evaluators (Figure 7.9). One should note however that even one evaluator is better than none!

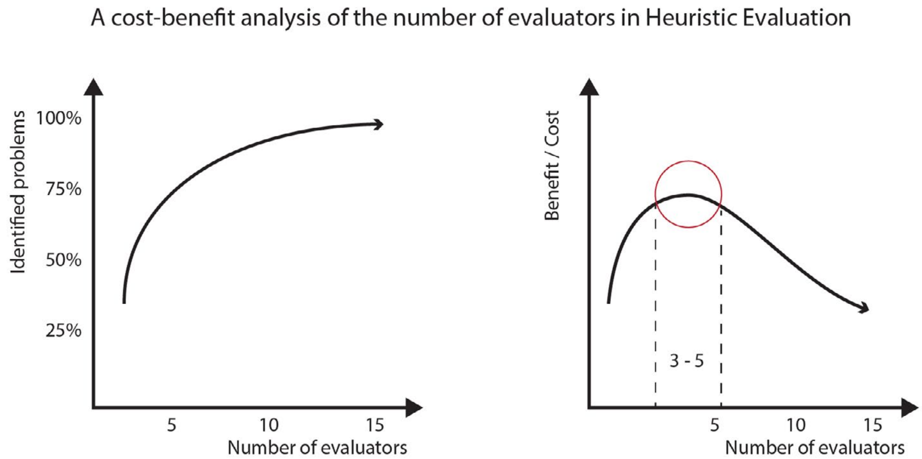


Figure 7.9. Jakob Nielsen's research indicates that five expert evaluators can help you discover about 75% of the usability issues. Beyond five, with every additional evaluator, the proportion of new usability issues will be much smaller and usually not worth the extra resources. © Jakob Nielsen and Nielsen Norman Group, Fair Use

Grading of the usability problems identified is based on a combination of the following criteria:

- Frequency of occurrence of the problem in the system.
- Impact on users, i.e., significance of the consequences on the users of the system (e.g., will they make significant errors? will they be delayed a lot? etc.) and how easily the problem can be solved.
- Persistence of the problem, i.e., is it a one-time problem that users can easily learn how to overcome, or will it bother them on an ongoing basis?

A well-known example of a usability problem with little persistence is the Microsoft® Windows® operating system shut down prompt in some older versions prior to Vista®. In those versions, in order to shut down the PC, users had to activate a button located at the bottom left of the screen named "Start" (Figure 7.10). While the wording was inconsistent with the function it served, in practice, users soon enough learned to overcome this inconsistency and it did not bother them at all.

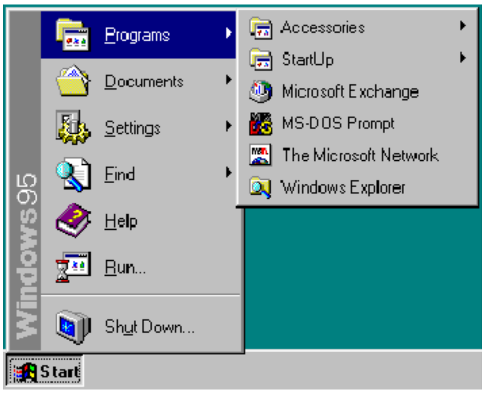


Figure 7.10. To shut-down a computer in Windows 95, users had to pass through the "Start" button; a rather unintuitive move. It turned out that the problem was not persistent since users would easily learn the path after performing it a couple of times.

To assess the severity of usability problems identified by evaluators, the following scale can be used:

0. I do not agree that this is a usability problem at all.
1. Cosmetic problem only: need not be fixed unless extra time is available on project.
2. Minor usability problem: fixing this should be given low priority.
3. Major usability problem: important to fix, so it should be given high priority.
4. Usability catastrophe: imperative to fix this before product can be released.

It should be noted that the Heuristic Evaluation can be performed at any stage of design –from the very initial to the most advanced– using either fully functional interface prototypes, or simple graphical representations (on paper or in any suitable software). However, it is considered more beneficial at the early stages of design and prior to user-testing, due to its simplicity, short duration and likelihood for quick fixes to the problems identified. An important drawback of the method is that it is not well suited for specialist or domain specific applications. These limitations occur because of the generic character of its heuristics.

To keep evaluators on the same track when conducting a Heuristic Evaluation, Jacob Nielsen proposed a list of Usability Heuristics which cover most of the generic usability problems on interfaces. Using such a list as common ground facilitates the assessment rendering the process briefer and more robust. Usually, not all Heuristics are taken into account when running the method. Instead, the design team selects those which considers more relevant or important for each specific project and informs the evaluators accordingly. The ten Usability Heuristics according to Jacob Nielsen (Nielsen, 1994a) are briefly presented below.

H1. Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable amount of time.

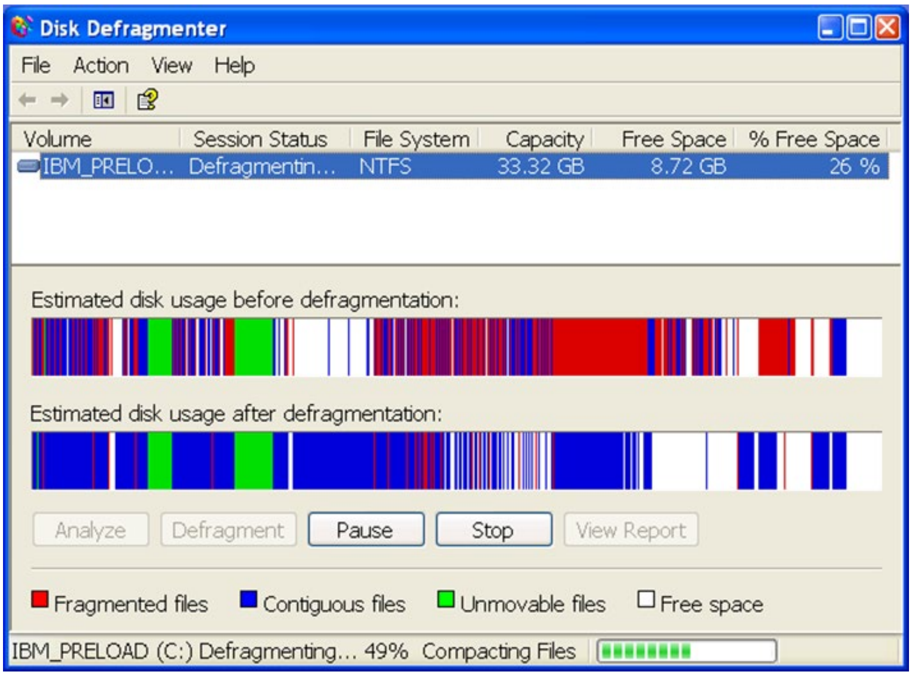


Figure 7.11. A disk defragmentation tool which does not only provide information on process status (e.g. remaining time), but also communicates a rough image of the disk's state before and after defragmentation.

Show feedback about:

- Time (how much time remaining?)
- Space (how much space remaining, e.g., in your webmail account)
- Change (e.g., the document has changed since your last save)
- Action (that the user is expected to perform some action next)
- System's actions (e.g., a notification message that an online transaction has been completed successfully, and the user should look for an email receipt)
- Completion (e.g., a long system process has completed successfully)

H2. Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms or technical jargon. Follow real-world conventions, making information appear in a natural and logical order.

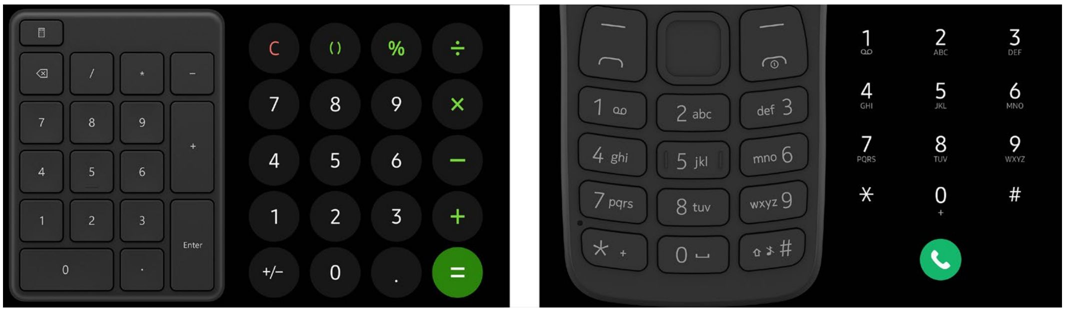


Figure 7.12. Established physical world conventions should be respected so that users migrating from the physical to the virtual interface will not be confused. Thus, a virtual calculator numerical pad uses the upward order of numbers (left) while a phone numerical pad uses the opposite (right).

How to be achieved:

- Ensure users can readily understand the terminology without having to go look up a word's definition.
- Never assume your understanding of words or concepts will match those of your users.
- User research will help you uncover your users' familiar terminology, as well as their mental models around important concepts.

H3. User control and freedom

Users often perform actions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue.

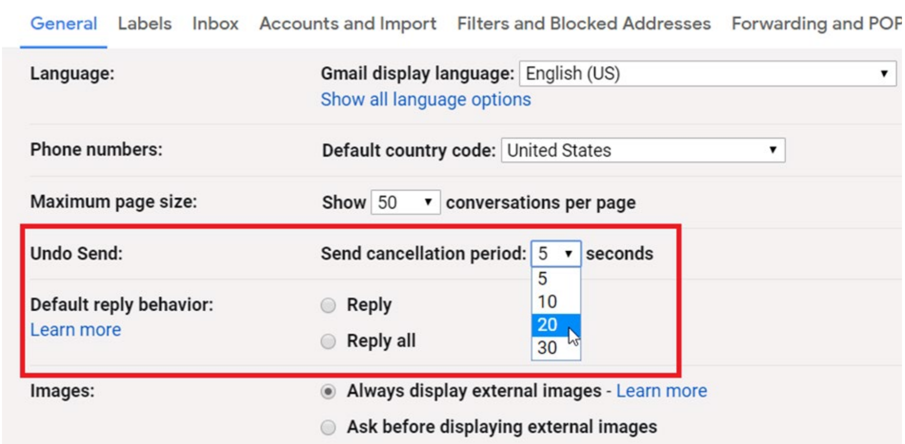


Figure 7.13. Allowing the user to set a delay period for e-mail sending after clicking on the "send" button, is a minor burden compared to the benefits of being able to cancel his action shortly after.

Specific guidelines:

- Support *undo* and *redo*.
- Also support freedom to explore possible system functions/actions.
- Make sure the exit is clearly labeled and discoverable.

H4. Consistency and standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform and industry conventions.

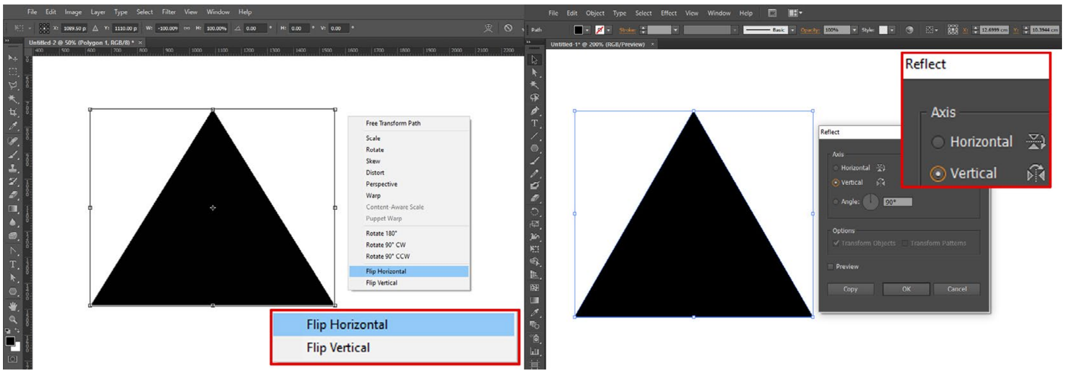


Figure 7.14. A quite different approach used for the same function. The "Flip Horizontal" function on the left matches the "Reflect Vertical" on the right interface, confusing the users that utilize both. The examples are taken from two graphic design software of the same company.

Specific guidelines:

- Improve learnability by maintaining both types of consistency: internal and external.
- Maintain consistency within a single product or a family of products (internal consistency).
- Follow established industry conventions (external consistency).

H5. Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

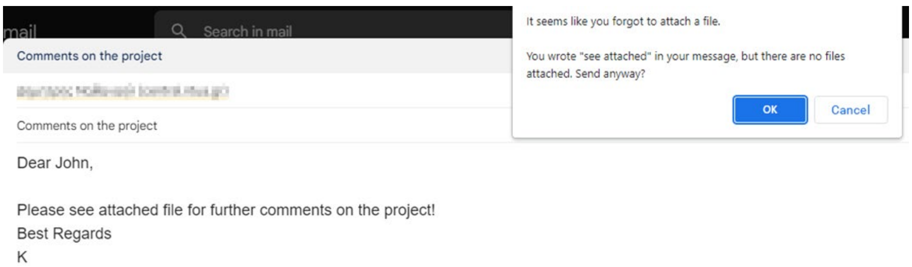


Figure 7.15. Simple but clever automated checks can prevent common user slips, such as forgetting to attach a file on an email.

Specific guidelines:

- Prioritize your effort: Prevent high-cost errors first, then little frustrations.
- Avoid slips by providing helpful constraints and good defaults.
- Prevent mistakes by removing memory burdens, supporting undo, and warning your users.

H6. Recognition rather than recall

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable when needed.

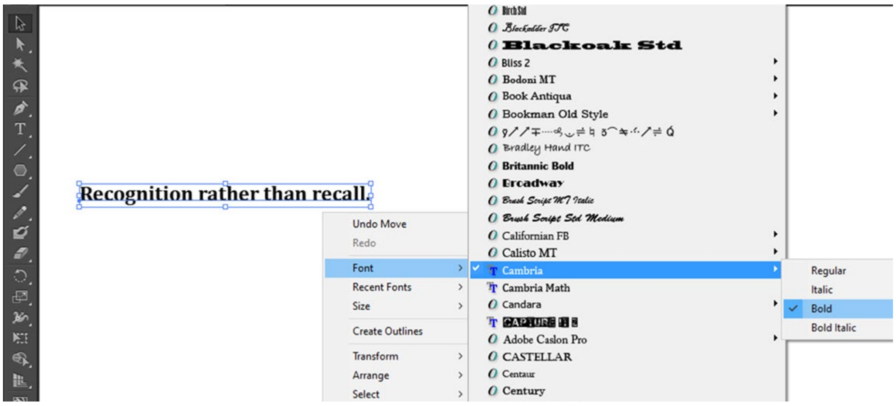


Figure 7.16. Recognizing the form of a font is much easier than having it memorize it only by its name.

Specific guidelines:

- Let people recognize information in the interface, rather than forcing them to remember it.
- Offer help when relevant, instead of giving users a long tutorial to memorize.
- Reduce the information that users have to remember.

H7. Flexibility and efficiency of use

Shortcuts —hidden from novice users— may speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Consider allowing users to tailor frequent actions.

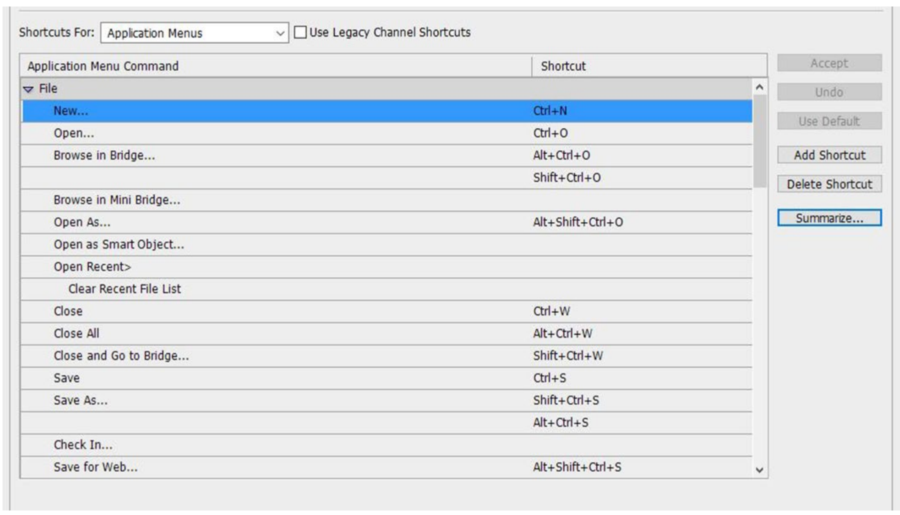


Figure 7.17. A good software not only provides shortcuts for the most frequent commands but also allows expert uses to customize them.

Specific guidelines:

- Provide accelerators like keyboard shortcuts and touch gestures.
- Provide personalization by tailoring content and functionality for individual users.
- Allow for customization, so users can make selections about how they want the system to work.

H8. Aesthetic and minimalist design

Interfaces should not contain information which is irrelevant or rarely needed. Every extra unit of information in an interface competes with the relevant units of information and diminishes their relative visibility.

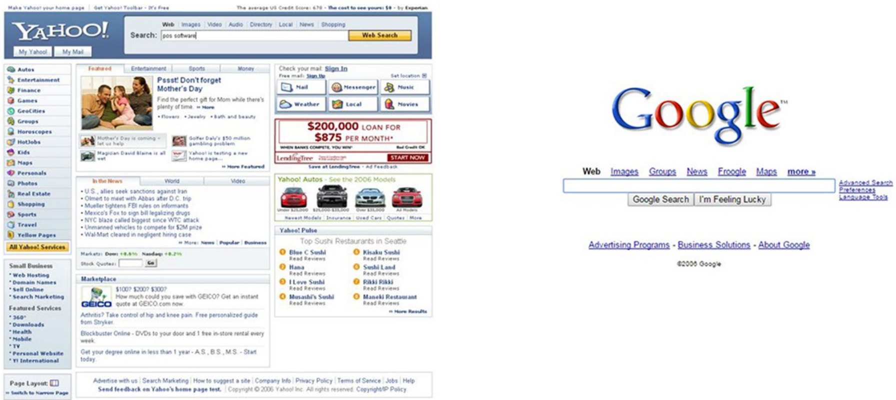


Figure 7.18. Back in 2006 Yahoo! and Google were the two main rival search engines. While Yahoo! had already had the largest user base, due to its wide range of services, Google quickly outperformed it, thanks partly to the straightforward and minimalistic design of its interface.

Specific guidelines:

- Keep the content and visual design features of the interface to the essentials.
- Do not let secondary elements distract users from the information they really need.
- Prioritize the content and features to support primary goals.

H9. Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no error codes), precisely indicate the problem, and constructively suggest a solution.

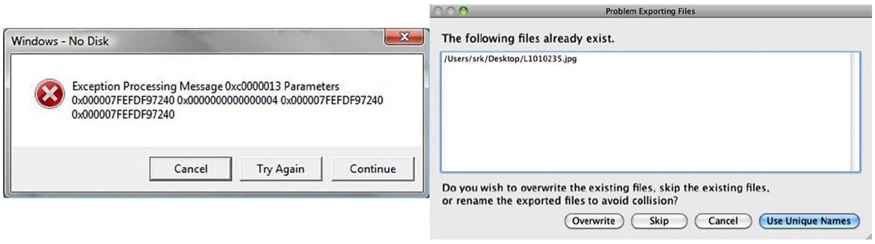


Figure 7.19. The error message on the right is so confusing that even the action options cannot help the user understand what must be done next. A good error message (right) not only explains the problem but also suggests meaningful actions.

Specific guidelines:

- Use conventional error message visuals, like bold, red text.
- Inform users what went wrong in terms they will understand —avoid technical jargon.
- Offer users a solution, like a shortcut that can solve the error immediately.

H10. Help and documentation

It is best of the system can be used without supporting written manuals. However, it may be necessary to provide some documentation. Any such information should be focused on the user's task, list concrete steps to be carried out, and not be too large.

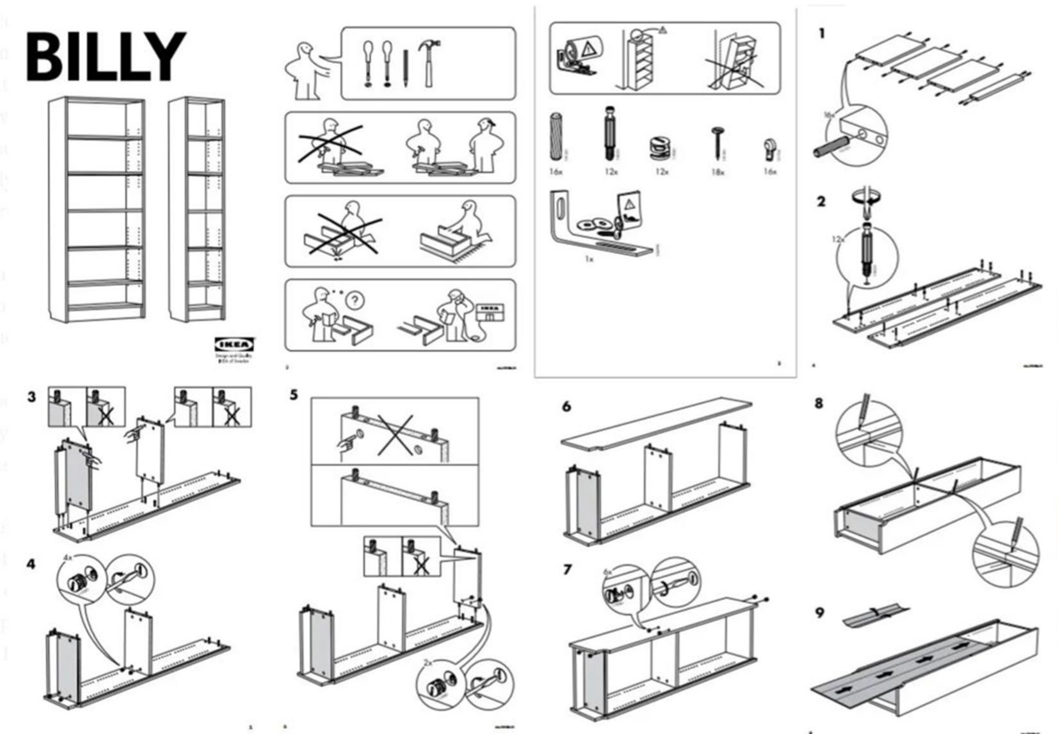


Figure 7.20. A well-known and award-winning instruction manual. Consisting of line figures and selecting only the most essential information to display, this manual can be readily followed by most adults without resorting to language.

Specific guidelines:

- Ensure that the help documentation is easy to search.
- Whenever possible, present the documentation in context right at the moment that the user requires it.
- Adapt documentation to different levels of user expertise.

Final notes on Heuristic Evaluation

Apart from formal evaluation purposes, the above list of heuristics can also be used as a "good practice" companion when designing interfaces, as it has proved its value through the years. Alternative lists of usability principles have also been proposed such as Gerhardt-Powals' (1996) cognitive engineering principles which take a more holistic approach to evaluation, Shneiderman's et al. (2016) eight golden rules of interface design and Weinschenk and Barker's (2000) classification of 20 heuristics. All the above can be used as good practice guidelines according to the needs of each project and the preferences of the design team.

7.3.3 Cognitive Walkthroughs

The Heuristic Evaluation method that we have just described, while ensuring the evaluation of the interface from many points of view, can be criticized of evaluating the system mainly "in a static manner", not considering in detail the sequence of actions that the user must perform in order to achieve specific goals, as well as without systematically taking into account the learnability of the system.

The Cognitive Walkthrough method (Polson et al. 1992) has been developed to address these weaknesses, by focusing on a dynamic exploration which simulates the way most users explore a new interface. The theoretical background of the method is the theory of "exploratory learning" by Polson & Lewis (1990). According to this theory, users familiarize themselves with an interface mainly by exploring it, rather than through instructions' manuals. More specifically, users typically start with a relatively vague formulation of the goals they want to achieve through the system. They then explore the user interface in order identify and perform the actions that are likely to enable the realization of their goals.

The justification proposed by Polson & Lewis (ibid.) for the preference of users for this way of learning over a nominal procedure through user manuals, is that users tend to invest only as much effort as it seems necessary to achieve their current goals. In other words, users prefer the immediate satisfaction of achieving proximal goals when performing a particular task and thereby gradually learning the artefact, rather than a "long-term" investment of learning it altogether. This approach ensures that the "cost" of gradual learning is in part proportional to the immediate benefit perceived by the user.

The approach we present on the method is slightly different than the one originally proposed by Polson et al. in 1992 and closer to that proposed by Rizzo et al. (1997). The difference lies on the questions asked at the 4^{th} step, which in our case follow more closely the consecutive steps of Norman's Action Cycle described in Chapter 3.

The evaluation process of the Cognitive Walkthrough method includes the following stages:

1. In a first phase, a team of evaluators is set up, which consists of usability experts that are familiar with the method. It is highly recommended that evaluators have not taken part in the design of the system which is about to be evaluated. Indeed, it is doubtful whether the designer of a system can apply the method alone with satisfactory results, given that s/he has an accurate and complete mental image of the system and, therefore, it is very difficult for him/her to adopt the perspective of future users.

2. Next, the evaluators proceed to develop use-case scenarios based on different personas relevant to the future users.

3. For each use-case scenario, the nominal actions to be taken by the users are exhaustively recorded, considering the design choices made in the interface.

4. For each use-case scenario and each user action, the evaluators, working as a team, ask the following questions (Figure 7.21):

    - Will users form the right goal, i.e., will they try to achieve the right result?
    - Will users associate the correct action with the result they are trying to achieve? Perhaps the correct button is visible, but will users identify it as the appropriate, and will they know how to engage with it?
    - Is the user able to perform the intended action? Can the action be easily performed in terms of physical power or dexterity of use?
    - After the action is performed, will users perceive and correctly interpret some feedback from the system? In other words, will the progress towards the goal be shown clearly?
    - Will users understand if their goal was achieved and if not, will they be able to understand why?

5. The application of the method is completed by grading/prioritizing the identified problems and searching for solutions.

Figure 7.21. Mapping the Cognitive Walkthrough questions to the different phases of the Action Cycle.

Associating the Cognitive Walkthrough with the Action Cycle, inevitably means that the evaluators should make good assumptions of the intentions, the knowledge, and the past experiences of the intended users. Indeed, the positive or negative answers to the five questions does not solely depend on the configuration of the interface per-se but is also highly dependent on the users' previous experiences. Therefore, while the method provides good results for the evaluation of interfaces aimed at intermittent use by the general public (e.g., ticket vending machines, e-shops, etc.), it is not considered suitable for the evaluation of domain specific applications, aimed at specialized users (e.g., design software, decision support systems). In these cases, iterative usability testing with the participation of a sample of future users is essential.

The duration of a cognitive-walkthrough session depends on the evaluators and the complexity of the tasks that are analyzed. It's generally possible to evaluate and document two medium complexity tasks in a 90-minute session. Like Heuristic Evaluation, Cognitive Walkthroughs not only identify usability problems but may also provide solutions. It is, therefore, considered a practical and effective method for assessing usability issues in walk-up and use, public domain interfaces.

7.4 User Testing

The cornerstone of User Centered Design is the testing of prototypes with prospective users. Unlike the expert evaluation methods described above, user testing methods are based on empirical evidence. They do not adhere to any pre-established theory of human activity or on the compliance of an interface with lists of heuristics. Instead, the effectiveness of the design is judged upon the user's actual engagement with the system, in realistic conditions, answering the most important question: "How does a user interact with a specific interface to accomplish a certain task?" Generally, user testing is more demanding in terms of duration, resources and technological equipment than expert evaluation methods, but provides more reliable data, especially at the latter phases of a design project. User testing can take many forms depending on the design phase but are all based on systematic observation of users performing predefined tasks. Observation data are subsequently analyzed through quantitative and qualitative metrics which may be compared to target values.

7.4.1 Setting up a user test

Conducting user testing requires an, at least partially, functional prototype and a sample of typical future users. The above also applies when running tests on paper prototypes through the Wizard of Oz technique which can be considered a form of primitive functional prototype.

The main steps of a user testing process are the following:

- Deciding on and writing of use-case scenarios that will be given to the participants. The number of scenarios to be written depends on the complexity of the system to be tested; scenarios should be judiciously chosen to encompass most critical aspects of a system without being exhaustive. Participant engagement is also important to this effect. The time requested from them should neither be too short, neither too long, so as to justify their contribution without being overly demanding in time and effort.
- Specification of the key quantitative and qualitative variables to be assessed (e.g., time to complete the task, number of errors, parts of the interface that make it difficult for users to proceed, etc.). This is one of the most challenging parts of the process since not only the variables must be specified but also their target values defined.
- Selection of the sample of test users. The number of participants is mostly depended on the phase of development, the complexity of the system and the available resources. As a rule of thumb, no less than three and no more than fifteen participants are needed to provide valid results. A common practice is to start with five and add more if no consistency is found in the results.
- Registration and briefing of test users. The users are informed about the procedure and their personal characteristics are registered (e.g., age, experience in using respective systems, etc.). Registering exhaustive user data that is not directly relevant to the task at hand should best be avoided. Even sex, educational level or age might not be relevant (e.g., as indicators of the expertise of someone). Try instead differentiating through seeking of the most relevant characteristics to the task at hand (e.g., prior experience with similar systems).

- Actual run of user tests. During testing, quantitative data such as the number of performed actions, the time taken to complete individual tasks, the wrong actions, etc., are recorded by appropriate means (see below). Depending on the testing protocol, participants are either allowed to ask questions when they find it hard to continue –in which case the hard point is recorded– or they are not allowed, at the risk of not being able to complete the task.
- Questionnaire administration. After testing, participants should typically fill out a questionnaire destined at collecting qualitative data (e.g., related to their experience while using the system, difficult points, misunderstandings, the aesthetics of the interface, or even suggestions for improvement).
- Data analysis. Finally, the analysis of the data collected during the process is carried out and the results are presented.

Data recording during user testing sessions can be done through various means, such as video recording of users, screen capture software, log-files or even eye-tracking recording. Depending on the needs the test apparatus can be either mobile or installed in a dedicated laboratory (Figure 7.22). Many laboratories feature specially designed double rooms, separated by a glass panel that offers visibility only in one direction. The participant is located in one room, while the moderators are in the other, allowing observation of the participant but preventing any observer effects (i.e., the disturbance of the participant by the act of being observed). Some practical advice on the testing process is provided below:

- Conduct a pilot test well in advance, with 2-3 pilot users to make sure your experimental equipment and recording apparatus works smoothly.
- Before each actual test run, conduct a practice activity to get your participants comfortable with the equipment.
- Remove distracting elements from the test area. Participants should not have to read or make note of anything during testing.
- If a moderator is present, s/he should sit next to and slightly behind the participant, so as not to encourage conversation.
- If the testing protocol allows questions from participants, the moderator should be brief and stereotypical in his/her answer(s).

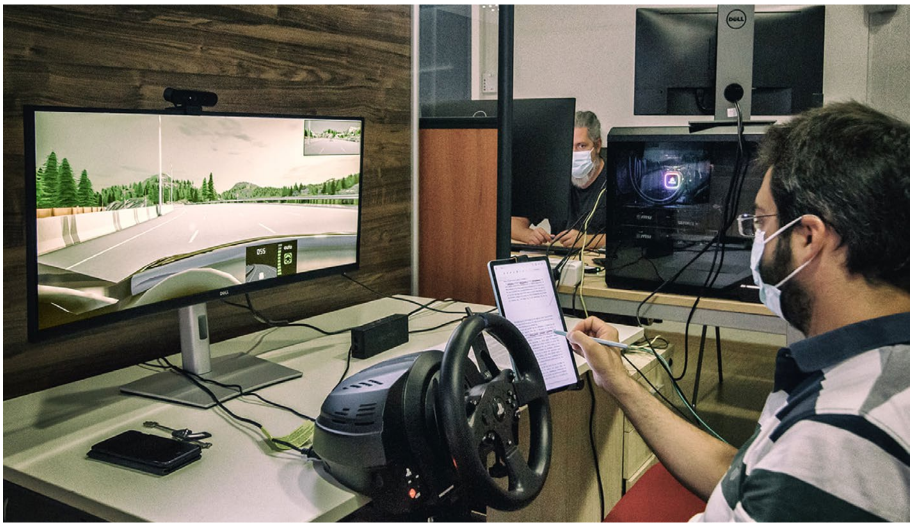


Figure 7.22. A user testing setup situated in a dedicated usability lab. A participant engages in a secondary task in automated driving mode on a driving simulator.

A note on eye-tracking tests

Especially for eye-tracking recordings, the setup is somewhat more sophisticated, (e.g., requiring calibration for each participant), and the collected data demand considerably more resources to be indexed and analyzed. Therefore, it is advised to resort to this technique only when it is considered important to answer specific perception related questions, such as: Which areas of an interface is the user inspecting to find a desired function? Which graphic elements attract the user's attention? In which order does a user scan an interface? Has a user looked at the desired function but did not identify it as the correct one?

Of the different eye-tracking technologies that have been discussed in Chapter 4, the most frequently used in usability testing is the monitor mounted, except for when there is need to evaluate a physical interface where the head mounted version is inevitable. The Nielsen Norman Group (https://www.nngroup.com) has outlined a number of useful good practices to follow when conducting an eye-tracking test.

7.4.2 Selection of metrics

The metrics to be assessed for each experimental task should be directly derived from the initial design requirements (often referred as "design brief") as their choice will have a non-negligible effect on the final design outcome. Depending on the particular design brief, various metrics may be needed, such as task completion time, success rate, number of errors, outcome quality, adopted strategy, etc. Nevertheless, the most common metrics used for non-critical and public domain systems are time to completion, number of actions to completion and success rate.

Target values for the metrics chosen will be set by the design team after completion of the prototypes to be tested. In a way, target values reflect the performance a design team hopes to achieve with a particular design solution. Therefore, nominal values based on optimal task flows may be used. For task completion time, a typical way to proceed is by timing the task executed by an expert user or by matching the optimal path and giving on both values a reasonable margin. For instance, if an expert user succeeds in two minutes and the shortest path consists of seven actions, then it would be reasonable to set the following goals for the user test:

- all users should be able to complete task X in less than three minutes,
- all users should be able to complete task X with less than twelve (12) actions.

Alternatively, target values can be set comparing the execution of the same task in similar products:

- all users should be able to complete task X faster in our platform than in the two rival platforms.

When the task consists of many actions or has many valid execution paths it is easier to count the number of errors

- 9/10 of users should be able to successfully complete task X with no more than two "undo" or "back" actions.

Metrics and target values can be set not only for the main task but also for any subtask of the process and for any segment of our users:

- type A users should be able to locate information X in less than one min.

In order to check the learnability of a system a number of consecutive trials with the same user may be appointed in a defined period of time.

- The time/errors should be reduced by x% while using the system for the third time in week.

A more nuanced metric is the percentage of trials where a user believes s/he has successfully completed a task while in reality, s/he has not. Such differentiation between the perceived and actual task accomplishment is quite critical as the user is unaware of the failure and will not seek remedial action.

In the case of eye-tracking, some additional metrics and associated values might be:

- function x should be identified as the correct one, on the first eye fixation,
- function x should be found no later than the first ten eye fixations on the website.

Many more variables can be tailored to each particular project, with target values that arise either from comparison with reference systems or from system requirements per-se. These could include “how many and what kind of assistance did the user get”, “which of the alternative paths was taken” or how these metrics were affected on a second or third session with the same user.

In addition to quantitative data, it is highly recommended to also collect qualitative data from users, attempting to grasp the overall experience of the user from a phenomenological perspective (e.g., satisfaction, excitement, anxiety, tediousness). This is best captured if users are prompted to write down their thoughts, ideas or experiences through open-ended inquiries such as:

- How did the execution of the task made you feel?
- Where there any points that you felt uncertain about any of your actions?
- Did you experience ambiguity when using the system at any stage?
- Would you consider the interface enjoyable to use?

It is essential to comprehend that human experience does not merely depend on task effectiveness but also on the feelings one gets through interacting with a system; if a system feels engaging, novel or enchanting, users will develop a positive attitude towards it, even if it renders them somewhat less effective in terms of task accomplishment. This is particularly important in the design of systems destined at discretionary users. A case in point is the considerable effort given to design enjoyable subtle feedback or transition animations in smartphones, the effect of which cannot be measured through objective measures, but only through the subjective experience of end users.

In conclusion, usability testing methods present several advantages, the main ones being:

- they provide reliable results linked to measurable objectives,
- they tend to identify more substantial problems than expert evaluation methods,
- they have strong convincing power towards all stakeholders (mainly due to the experimental approach and the use of quantitative measurements).

However, they also present some drawbacks, the main ones being the following:

- they require considerably more resources than expert evaluation methods,
- they can be reliably applied only after a functional prototype has been developed,

- the experimental conditions may affect the participants behavior,
- they require a relatively large number of typical users, who are not always easy to recruit,
- they do not provide direct design solutions to the identified problems.

7.4.3 Experimental design

The term “experiment” while is commonly associated with any kind of trial of a new condition in order to see what “will happen”, is, in fact, a scientific method which is governed by specific rules aiming at testing the validity of a hypothesis. More specifically, an experiment is an empirical test designed to answer the question of whether there is causal relationship between two or more variables. In other words, an experiment tests whether the changes in one variable –called independent variable–, have an impact on another variable –called dependent variable– (Jennings, 2005; Jhangiani et al., 2019). Establishment of causality is the main objective of an experiment, so a straightforward question to be tested could be “*How is time-to-task completion affected when changing the position of a call-to-action button?*” The element that presumably causes the measured effect is the independent variable (i.e., the position of a call-to-action button). The dependent variables are the measurements that would reveal any causality (i.e., the time-to task completion, the number of mistakes, the number of participants that pushed a certain button etc.), when manipulating an independent one. Sometimes the causality is not directly hypothesized but is implicit in questions like “*Why is this design more effective than another?*”, “*Which elements cause its effectiveness?*”) or “*Are users less prone to errors when using this interface?*”. As stated above, the independent variables are hypothesized as causes in the sense that a change in the independent variable (or “the cause”) influences a change in the dependent variable (or “the effect”). Note that such linear causality in the proper sense is difficult to verify. For instance, two variables may be correlated suggesting interdependency between them, however, if no theoretical explanation can be made for how one causes the effect to the other, then the criterion for causality is not satisfied. Without going into the intricate details of causality, let us adhere to the following basic criteria: (i) there must be an empirical association between the variables, (ii) there must be an appropriate time order, i.e., the “cause” must take place before the “effect”, and (iii) the association between the variables must not be due to a third, or confounding, variable.

In an experimental setting, in order to claim that the causality is due to a certain element, all other features of the experimental environment that might affect the dependent values should be kept constant. Otherwise, the experiment can be “biased” by external elements and provide false results. In experiments studying human behavior, things can get quite complicated. First and foremost, human behavior cannot be considered as purely reactive, i.e., affected solely and unilaterally by external factors.

Thus, even in cases where external conditions are fully controlled, causal relationships between a certain element and human “response” can only be established statistically (i.e., over multiple measurements across subjects or over time).

In order to eliminate external disturbances, as far as possible, experiments of human behavior are usually conducted in laboratory conditions (e.g., usability labs). However, laboratory conditions, although strictly controlled, also affect participants’ behavior due to the unnaturalness of the setting. Therefore, often, it is preferred to run user tests in less controlled environments trading formal experimental validity for a more naturalistic feeling. All such tests that do not follow the strict rules of experimentation cannot be properly considered as experiments but are instead called “quasi-experiments”.

The quality of experiments in social sciences is generally being assessed through their **Reliability** and **Validity**.

Reliability refers to how consistently a method measures something. If the same result can be consistently produced using the same methods under the same circumstances, the measurement is considered reliable.

Validity refers to how accurately a method measures what it is intended to measure. For example, “how valid is the IQ test in measuring intelligence?” or “how valid is the performance of a participant in a laboratory compared to real-world conditions?” If the research has high validity, this means it produces results that correspond to real properties, characteristics, and variations in the physical or social world. (Middleton 2022).

While reliability is rather straightforward to grasp, validity, in a social study, can only be assessed in regard to how it is used, and what interpretations are given to the scores for particular groups of people. A test that may be valid (i.e., yield useful and accurate information) for one group or in one setting may be completely invalid with other people or in other situations (Steiner 2006).

To ensure validity we must:

- Choose appropriate methods of measurement.
- Use appropriate sampling methods to select our subjects in terms of number and characteristics.
- Run the appropriate number of sessions to reach statistical significance.

To ensure reliability we must:

- Apply our methods consistently.
- Standardize the conditions of our research.
- Include all these data in the report, so that other researchers can replicate the same procedure and comparable results.

An experiment may begin with a hypothesis of whether a condition is true or false (i.e., "*The new interface design is more effective than the previous one*"). The testing of this hypothesis is actually a comparison between two conditions of the independent variable "interface design" (old vs new). The design that serves as comparison for a new one is called the *Control condition*, while the new one *Variation A*. In some cases, more variations (*Variation C, Variation D,* etc.) can be compared to the Control condition as well as between them.

In the Selection of Metrics subsection, we have talked about target or reference values with which we typically compare our new design. If no reference values for comparison have been established (e.g., from a rival design or from the old one), we are compelled to run the test for both old (control) and new designs. To do so we must decide how to allocate our participants to the designs to be tested. Will all of them run a particular task in both designs or will they be split so that half of them interact with the control and the other half with the new one? Both methods have pros and cons.

Let's present an illustrative example: consider a Usability Test of two slightly different virtual keyboards (A and B) for tablets and suppose we have recruited six participants. To test the keyboard's usability, participants must type a given text while we will be measuring the time it took them to accomplish the task and the errors they made. If all participants are tested on both keyboards, then this is called a "*within-subjects*" experiment, because the measurements from each condition are taken from the performance of all participants (Figure 7.23). In this case we would have data from all six participants on both designs and the comparison will give us rich and directly comparable data. But since the task is the same for both keyboards, it is inevitable that participants will be better in the second trial due to a *learning effect.* To overcome this phenomenon, we can assign half of the participants to begin with keyboard A and the other half with keyboard B. To further reduce the learning effect, we may use alternative texts, with the same number of letters and similar complexity, for the first and the second trial (Figure 7.24).

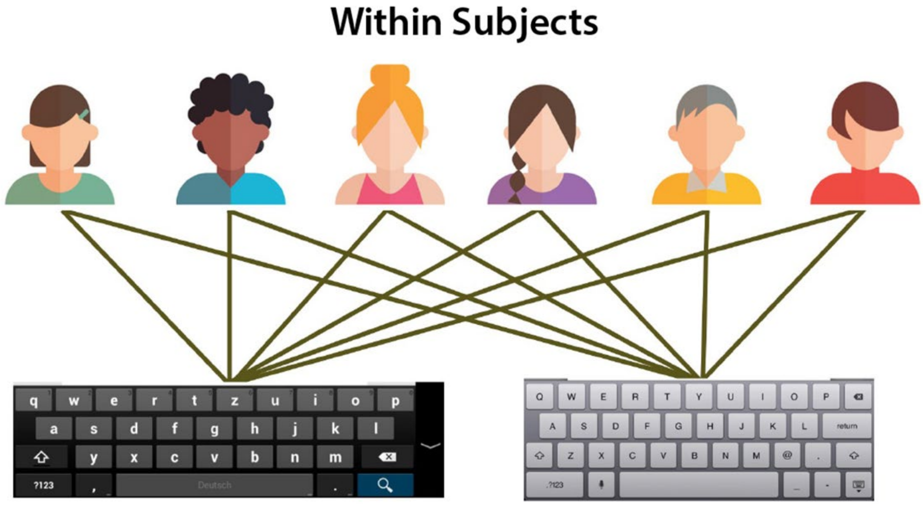


Figure 7.23. Within subjects experimental design.

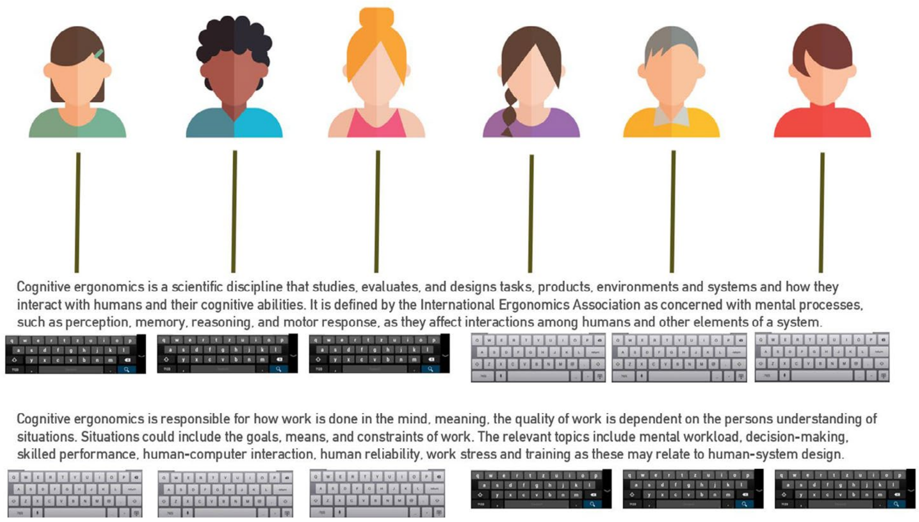


Figure 7.24 Counterbalancing learning effect by alternating test order trial and typing text.

The obvious way to entirely avoid the learning effect is to split the participants' sample in half, and assign each group to one of the two designs. This method is called "*between-subjects*" as the comparison is taking place between the performances of the two groups (Figure 7.25). This method, however, significantly reduces our trial sample, so it requires the recruitment of more participants and more resources. Even so, in this case we have to deal the issue of equivalence between the two groups. Since each user will only be tested in one condition, how can we know their performance isn't mainly due to some exceptional typing expertise rather than on the design being tested? What's more if by chance the most skillful participants are in the same group, the comparison between the two keyboards would be jeopardized.

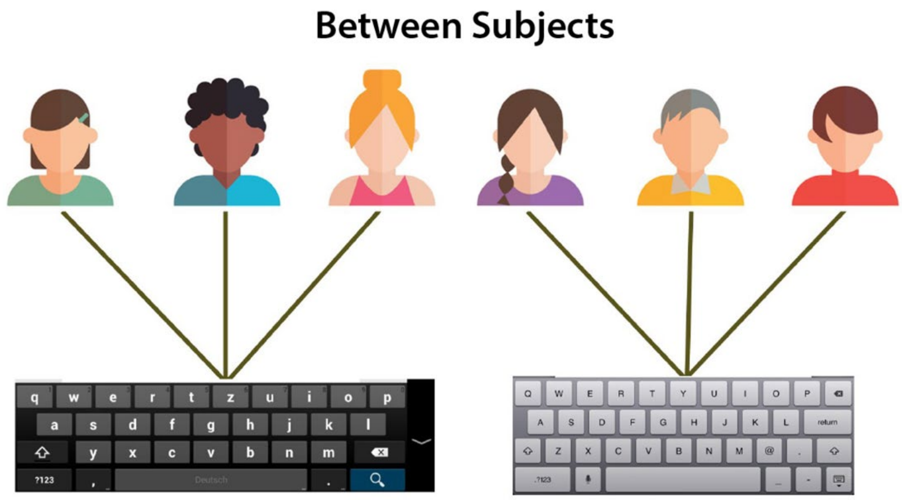


Figure 7.25. Between subjects experimental design.

To eliminate this possibility, we can run a preliminary ranking test, for instance, by giving our participants a similar text and asking them to type it on another similar keyboard. We can then rank the participants by completion time (or number of errors) and allocate them to the two conditions according to the ranking test. A common way to proceed is by creating consecutive participant couples, beginning from the top to the bottom of the ranking list. Then, the two members of each couple are assigned to different groups alternatively, so that, for example, the most skilled subject is assigned to group A, the second is assigned to group B, the third to group B, the fourth to group A, etc. In this way, two similarly skilled groups are created, and we can, then, run the main experiment. (Figure 7.26)

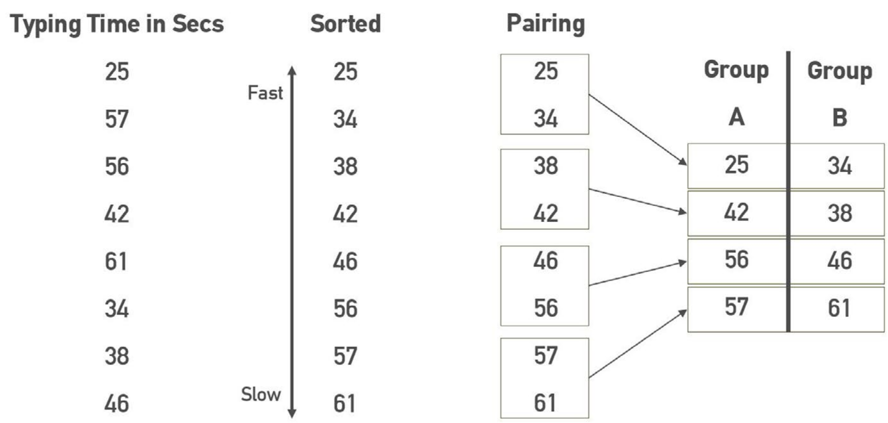


Figure 7.26 Results of a preliminary ranking test and pairing of participants in two similarly skilled groups.

However, depending on the complexity of the task, such ranking might not always be easy to achieve through preliminary testing. In such cases, we can instead rank the participants in a normative manner. For instance, we may infer driving skill by asking questions such as "*how many years do you hold your driving license*?" and "*how many km do you drive each year?*".

Such ranking data are, off course, less reliable than preliminary tests, but they can still be used if empirical ranking data are impractical to obtain.

7.4.4 A word of caution

As noted earlier, human behavior cannot solely and unilaterally be determined by external factors as it inescapably entails internal mechanisms such as interpretation, emotion and motivation. Thus, even in the controlled conditions of an experiment, human behavior may be significantly affected by purely subjective factors. Here is a famous example. In a series of studies that took place between 1924 and 1932 at the Hawthorne Works (a Western Electric factory outside Chicago) the goal was to examine the effect of lighting levels to the productivity of the workers. The study showed that a slight increase in illumination levels had a positive effect on workers' productivity. In subsequent increases in illumination, worker productivity kept rising. After several attempts trying to define the optimal illumination levels, the researchers noticed, to their surprise, that productivity kept rising even when decreasing the illumination! In fact, workers' productivity seemed to improve right after changes were made and ultimately returned to previous levels when the study finally ended. It was, therefore, suggested that the productivity gain was mainly due to the motivational effect on the workers, of the attention being given to them (Cox, Erika 2000). This phenomenon was named "the Hawthorne effect" after the factory in which the study took place. There are various ways to tackle this effect, such as concealing the purpose of the research from participants or deliberately lead them to think they are being evaluated in different aspects of the task than they are actually measured. This in turn raises ethics issues that must be addressed accordingly.

Ethics is nowadays considered of high importance when working with human subjects and this is partly reflected in the recent legislation on General Data Protection Regulation (GDPR) concerning matters of privacy. Apart from that, issues on autonomy, freedom of choice and informed consent constitute the pillars of people's rights and dignity as human beings. According to Jhangiani et al. (2019), in all relevant user studies, the researchers must:

- Know and accept their ethical responsibilities.
- Identify and minimize risks for participants.
- Identify and minimize deception.
- Weigh the risks against the benefits.
- Create informed consent and debriefing procedures.
- Get institutional approval.
- Stay consistent to the protocol, during and after the end of the research.

The most accepted ethics code today is the American Psychological Association's *Ethical Principles of Psychologists and Code of Conduct* (also known as the APA Ethics Code) which can be found here: https://www.apa.org/ethics/code/index

7.5 Post-launch monitoring and A/B Testing

Design work does not quite end when a system is launched. Indeed, it should be, and often is, an ongoing process that continues through the system's lifecycle. System use in real conditions by its actual users is in-fact the ultimate test. Continuous feedback from actual use will help the design team to learn and gradually improve it.

The most cost-efficient way to monitor actual use is thorough quantitative analytics tools (e.g. Google analytics ™). Metrics provided by an analytics tool (clicks, visit-time, navigation paths, search queries, etc.) can be used to understand how users actually interact with the system (Figure 7.27). Metrics can also uncover bugs, unforeseen behaviors or user requests that were not captured in user testing. Online surveys can also be used to gather qualitative data, although these should be employed sparingly and be clearly targeted to specific issues.

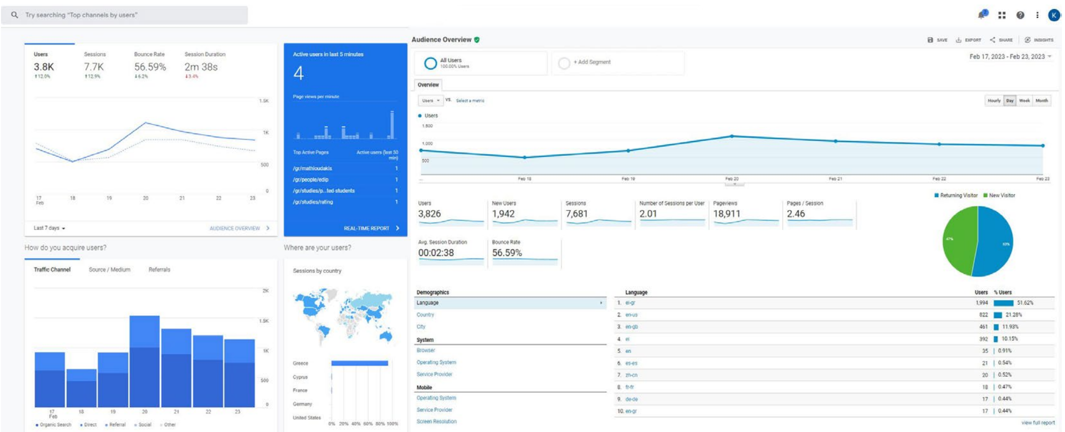


Figure 7.27. A real-time monitoring of a web site which provides metrics such as the number of active users, the bounce rate, the average session duration and many more.

System monitoring will eventually call for system rectifications or improvements through re-design. However, system alterations during use are often risky, as expert users are already familiar to a particular interface and often averse to changes. Even small alterations in the arrangement of controls might be frustrating to frequent uses as it disrupts their ability to work on a skill-based level (see SRK model) demanding to recalibrate their established subliminal routines, thus making them less effective for a given period of time. Occasional users, on the other hand, are less sensitive to changes but it is not always clear whether an alternative design will suit them better than the original. To check if small design changes positively affect certain aspects of a system-in-use we often use a particular method called A/B testing (or s*plit-run testing*).

Understanding the rules of experimentation is crucial for conducting A/B testing, as this is a purely quantitative method which relies on big data. Conducted only on finalized products that are already online, A/B testing consists of a randomized experiment to compare two versions (A-control and B) of a single variable in terms of effectiveness towards a predefined goal. On a news webpage, for example, we might need to check how the repositioning of the “subscribe” call-to-action button will affect the number of new subscriptions over the previous design. To ensure that the subscriptions are not affected by external factors, both designs must be tested simultaneously on two different groups of users (called Control and Treatment), making the method a *between subjects quasi-experiment*. The division of subjects into groups can be achieved through various means (e.g., their Internet Protocol Addresses) in such a way as to ensure group equivalence in terms of demographics, operating system, screen resolution, etc. Users are tested while naturally interacting with the system and are most often unaware of taking part in a test, so as to avoid affecting their behavior; in most cases users perceive such changes as ordinary updates. The sampling duration depends on the time necessary to achieve statistical significance on either accepting or rejecting the null hypothesis. In most cases A/B tests last around two weeks.

The main advantage of A/B testing is that it records actual user’s behavior since it is conducted without their knowledge, through their own hardware and in their natural environment, eliminating all factors that affect data validity. Also, it can provide extremely reliable data since it can incorporate large numbers of trials with no additional cost. If set up correctly, the data can be directly analyzed and provide a definite answer to the initial hypothesis. On the downside, A/B testing can only examine a single change at a time since changing more than one element simultaneously could produce complex interactions rendering the results inconclusive. In addition, the method, being purely quantitative, can only be applied to large numbers of users, and thus, is inadequate for the early stages of design or for functional prototype testing. Lastly, there is always concern whether the change being measured through A/B testing will indeed have a permanent effect on user’s behavior, since it only measures short term outcomes.

References

Cathleen Wharton, John Reiman, Clayton Lewis, Peter Polson. (1994). The cognitive walkthrough: A practitioner's guide. In Jakob Nielsen, Robert L. Mack (ed.) *Usability Inspection Methods,* John Wiley & Sons Inc, New York, New NY. https://dl.acm.org/doi/book/10.5555/189200

Clayton Lewis, Peter Polson, Cathleen Wharton, John Reiman (1990). Testing a Walkthrough Methodology for Theory-Based Design of Walk-Up-and-Use Interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*, April 1-5 1990, Seattle Washington USA, Association for Computing Machinery, New York, NY, 235-242, https://dl.acm.org/doi/10.1145/97243.97279

Cox, E., (2000). *Psychology for AS Level*. Oxford: Oxford University Press. p. 158. ISBN 0198328249.

David L. Streiner, Geoffrey R. Norman, "Precision" and "Accuracy": Two Terms That Are Neither, *Journal of Clinical Epidemiology*, Volume 59, Issue 4, 2006, Pages 327-330, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2005.09.005

Gayle R. Jennings, (2005) Business, Social Science Methods Used in, Editor(s): Kimberly Kempf-Leonard, *Encyclopedia of Social Measurement*, Elsevier, Pages 219-230, ISBN 9780123693983,

Gerhardt-Powals, Jill (1996). "Cognitive engineering principles for enhancing human – computer performance". *International Journal of Human-Computer Interaction*. 8 (2): 189–211. https://doi.org/10447319609526147

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Vermeer, L., (2019). *Top Challenges from the first Practical Online Controlled Experiments Summit*. ACM SIGKDD Explorations Newsletter. 21. 20-35.

Hodgson P. (2014) *How to Experiment*. User Focus.co.uk. https://www.userfocus.co.uk/

International Organization for Standardization. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts* (ISO Standard No. 9241-11:2018). https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Jhangiani, R., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research Methods in Psychology (4th ed.).* Surrey, BC: Kwantlen Polytechnic University. Retrieved from https://kpu.pressbooks.pub/psychmethods4e/

Middleton, F. (2022, October 10). *Reliability vs. Validity in Research | Difference, Types and Examples*. Scribbr. Retrieved November 21, 2022, from https://www.scribbr.com/methodology/reliability-vs-validity/

Molich, R., and Nielsen, J. (1990). Improving a human-computer dialogue, *Communications of the ACM* 33, 3 (March), 338-348.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, NY, USA, 152–158.

Nielsen, J. (1994b). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), *Usability Inspection Methods*, John Wiley & Sons, New York, NY.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People* - CHI '90. http://doi.org/10.1145/97243.97281

Rizzo, Antonio & Marchigiani, Enrica & Andreadis, Alessandro. (1997). The AVANTI Project: Prototyping and Evaluation with a Cognitive Walkthrough Based on the Norman's Model of Action. *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques,* DIS. 305-309. http://doi.org/1 10.1145/263552.263629

Shneiderman, B., Plaisant, C., Cohen, M. S., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the user interface: strategies for effective human-computer interaction*. Pearson.

Weinschenk, S. & Barker, D. (2000). *Designing Effective Speech Interfaces*. London:Wiley

Chapter 08: Interface Design

Chapter 08: Interface Design

Chapter Summary

The chapter discusses issues of detailed interface design. Before an interface or a tool (in either electronic or physical form) takes its final form, various details of how to present information and of how to design controls need to be decided upon. These detailed design issues play their own distinctive role in the quality of the final product. To tackle these issues, a number of key notions on human visual scanning, perception and interpretation are provided, starting from visual perception principles through semantics and sociocultural aspects. Next, several guidelines for visual design are presented to make the designs more comprehensible and effective, followed by design principles for displays and controls, with particular emphasis on visual aspects. In essence, the following issues are addressed: the most appropriate way of presenting information depending on the task at hand, optimizing information content, the pros and cons of various types of control and general considerations when designing display / control ensembles. Throughout the chapter, typical real-world examples are provided aiming to help the reader in getting a better grasp of the connection among physical appearance, tangibility and cognition.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of the previous chapters.

8.1 Visual design principles

The constant movements of the human eyes when inspecting a visual scene may be largely involuntary but are not random. The object's contours, the boundaries between surfaces with different colors or light intensities, as well as any change in the field of view, attract our vision towards them, so that detailed visual information can be collected. This constant eye movement may explain how "we see / perceive" complex or even impossible objects, such as those shown in Figure 8.1.

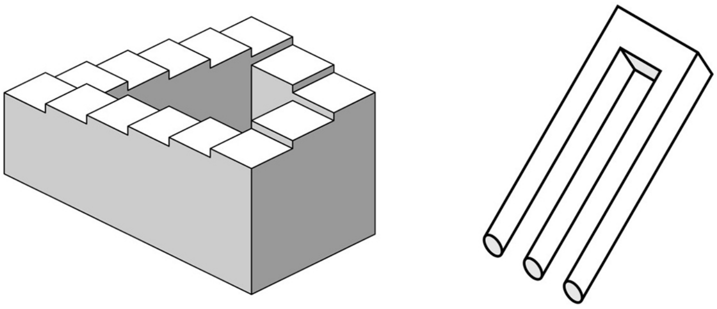


Figure 8.1. Constant eye movement makes us perceive even impossible objects.

8.1.1 Gestalt perception basics

Gestalt psychology was a school of psychology that emerged in the early twentieth century in Austria and Germany, mainly as a theory of perception. Gestalt psychologists emphasized that organisms perceive entire patterns or configurations, not merely individual components. The view is sometimes summarized using the phrase “the whole is more than the sum of its parts”. Gestalt principles, such as proximity, similarity, closure, and continuity, determine how humans perceive visual stimuli in connection with different objects and environments. Although Gestalt theory is largely descriptive with no explanatory value, and not sufficiently supported by quantitative findings, it has proved very useful on several visual design fields, such as user interface design and cartography. In fact, the Gestalt principles constitute a universal way of visual meaning creation, and thus by understanding them we can create visual content that is easy to perceive and interpret.

Proximity

Elements which are close to each other are perceived as forming a group (Figure 8.2).

Figure 8.2. Elements close to each other are perceived as groups despite their differences in color or shape.

Similarity

Elements within an assortment of objects are perceptually grouped together if they are similar to each other (Figure 8.3).

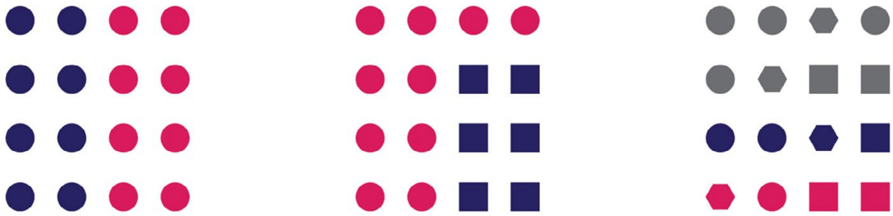


Figure 8.3. Similarity among elements gives the impression that they belong to the same group.

Closure

Human perception tends to recognize objects as complete rather than focusing on the gaps that the object might contain (Figure 8.4).

Figure 8.4. Left: we tend to perceive a white pentagon on top of five circles instead of five major sectors. Right: we tend to perceive this shape as an incomplete circle rather than as two separate arcs.

Symmetry

Human perception tends to see symmetrical objects as forming a group (as in similar ones). The symmetry law often prevails in grouping salience over the proximity one (Figure 8.5).

Figure 8.5. Elements that are symmetrical to each other tend to be perceived as a unified group.

Common Fate

Objects are perceived as part of the same group when moving along the same direction. This law is better understood in scenes with motion and used widely in motion graphics and cinematography (Figure 8.6).

Figure 8.6. In an equivalent animated figure, the circles that move towards the same direction would be perceived as belonging to the same group.

Continuity

Elements arranged smoothly in a continuum tend to be grouped together, forming integrated perceptual wholes or high-level entities (Figure 8.7).

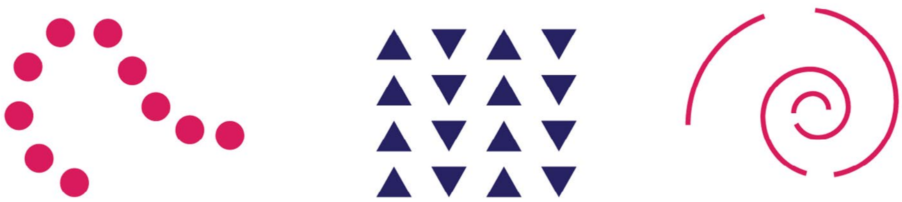


Figure 8.7. Elements that seem to follow a certain path or being constitutive of a bigger entity are seen as a group.

Past Experience

Under certain circumstances, visual stimuli are categorized according to past experience. Besides the proper visual characteristics of the objects inside the visual field, the scene inspection is also influenced by prior knowledge, experience, or the task at hand. This fact can explain why a sentence is easier to read when typed in lower case letters than in uppercase ones or why we can easily recognize a word only by looking at its upper half (Figure 8.8). The idea of how users' experiences affect the perception and interpretation of visual stimuli will be discussed in depth later.

Figure 8.8. One can recognize the words in the phrase above by only looking at its upper part, based on her/his past experience of reading.

8.2 Practical Guidelines for visual design

Having decided in broad terms on the semantic, spatial, and sequential configuration of information, one needs to consider specific graphical implementations of the above on the available space. Here, the visual integration principles, derived from Gestalt theory, are relevant.

8.2.1 Minimalism

The main assertion of Gestalt theory is that the human brain is hardwired to extract or even invent structure out of incomplete external stimuli. This property is, actually, a manifestation of the human brain's innate tendency towards trying to make sense of the world, i.e., meaning making (see Chapter 2 on signs and mental models). Exploiting this innate characteristic of the human brain, a designer can convey maximum meaning out of relatively limited visual stimuli (Figures 8.9, 8.10).

Figure 8.9. The figure above is recognized as the word "THE CAT" despite employing the same obscure symbol for both H and A, manifesting both the law of closure and that of past experience.

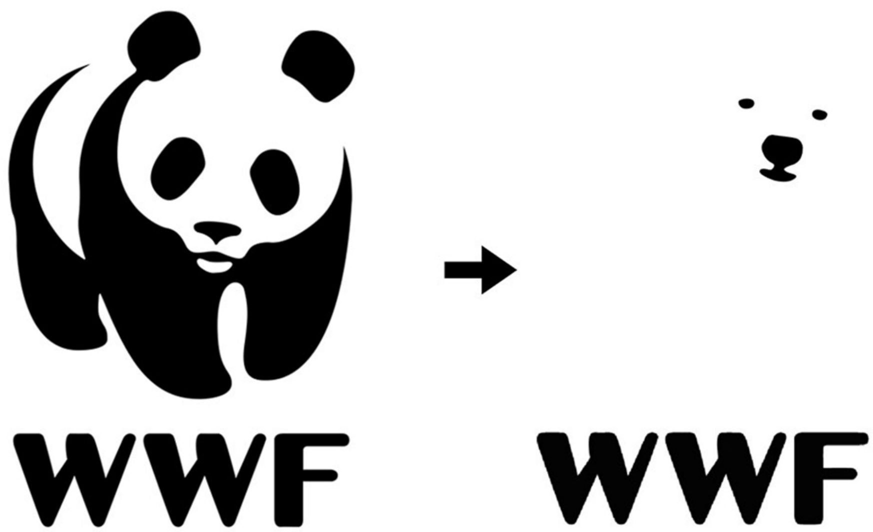


Figure 8.10. Humans will instantly recognize a panda and a polar bear out of quite incomplete and abstract forms (The "disappearing arctic bear graphic is a public proposal for a new World Wildlife Fund logo by Grey London design agency).

8.2.2 Visual structure

As mentioned above, humans tend to search for meaning in any visual configuration, irrespective of it being purposely designed or not. For instance, a text page with arbitrarily scattered paragraphs will inevitably provoke a meaning making effort in the human viewer, even if it remains inconclusive. This explains why it is standard practice in graphic design to use a grid for organizing visual content. The use of the grid attenuates noise and ambiguity, by doing away with unintended probes for meaning. A non-visible grid will achieve the same effect with even less visual clutter (Figures 8,11, 8.12 and 8.13).

Figure 8.11. The same information in three visual configurations: with no grid (left), with a visible grid (center) and with invisible grid (right). The invisible grid configuration has minimum visual noise, doing away with both random and redundant information.

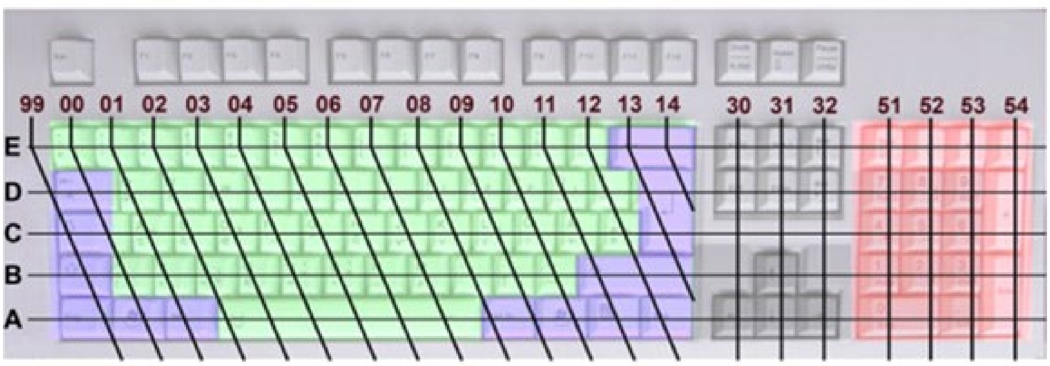


Figure 8.12. Example of grid use on a physical interface.

Figure 8.13. The reason this parking pay center seems confusing is that it lacks the basic Gestalt principles of proximity and similarity. Moreover, the elements instead of being organized on a grid, are scattered randomly on the interface, causing puzzlement even to a user who is acquainted with such a service.

8.2.3 Visual Intensity

Graphic design details (i.e., such as the relative size and layout of individual elements, the color palette or shapes used) will have a profound effect on the visual intensity of the presented information. Visual intensity refers to the overall energy in an image, energy created from the variation or contrast between its visual components.

The human visual system is attracted by contrast. Because we are exposed to so much information each day, our brains have developed ways to efficiently process all this information by ignoring areas of small visual energy. This makes that, irrespective of our will, our eyes are attracted by elements of high contrast relative to their surroundings. As a result, when designing a visual interface, alongside a grid, one must specify a visual intensity scaling in the various elements (Figure 8.14). Correct visual intensity scaling becomes critical in complex interfaces, consisting of a large number of elements. In such cases, elements with comparable visual intensity tend to fight each other in attracting the user's eye, resulting in decreased overall intelligibility of the interface.

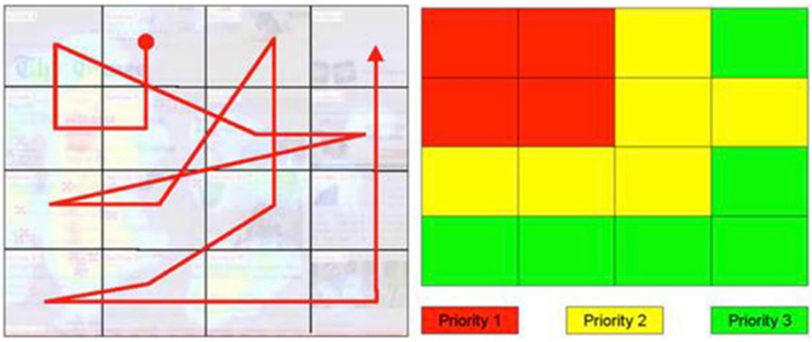


Figure 8.14. Typical visual scan path when viewing a computer screen composed of elements of varying visual intensity.

8.2.4 Optimal condensation of meaning

Consider the three alternative ways to convey the information that four out of seven people agree on a certain topic, presented in Figure 8.15.

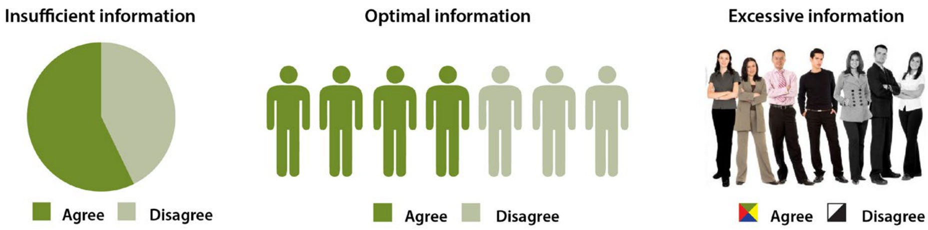


Figure 8.15. Three alternative ways of presenting the same information. The one in the middle outperforms the other two in terms of meaning condensation and clarity.

While the design on the left is considered insufficient in terms of the actual number and nature of the sample, the one on the right has so much "informational noise" that distracts the viewer from the main point. The middle one is considered to have an optimum amount of information which manages to inform without confusing.

Putting informational noise aside, even accuracy of information is not always desirable. For instance, take a look at a geographically accurate map of the London Underground in Figure 8.16. The lines' density at the center of the city along with some steep corners and overlaps make the map confusing and copious to consult for train commuting. In contrast the official London underground map in Figure 8.17 looks much more effective for commuting without being geographically accurate. The idea of the official design is to maintain a consistent grid and only use line directions of 0°, 45° and 90° degrees, distorting actual geographical distances and direction among stations but enhancing clarity for commuting planning. In a narrower version that is found inside the train carriages, the map of a certain line has been entirely deprived of any geographical accuracy (aka "straight-line diagram") as the main concern of the passengers is not the actual position of the stations but their number and the upcoming junctions with other lines (Figure 8.18).

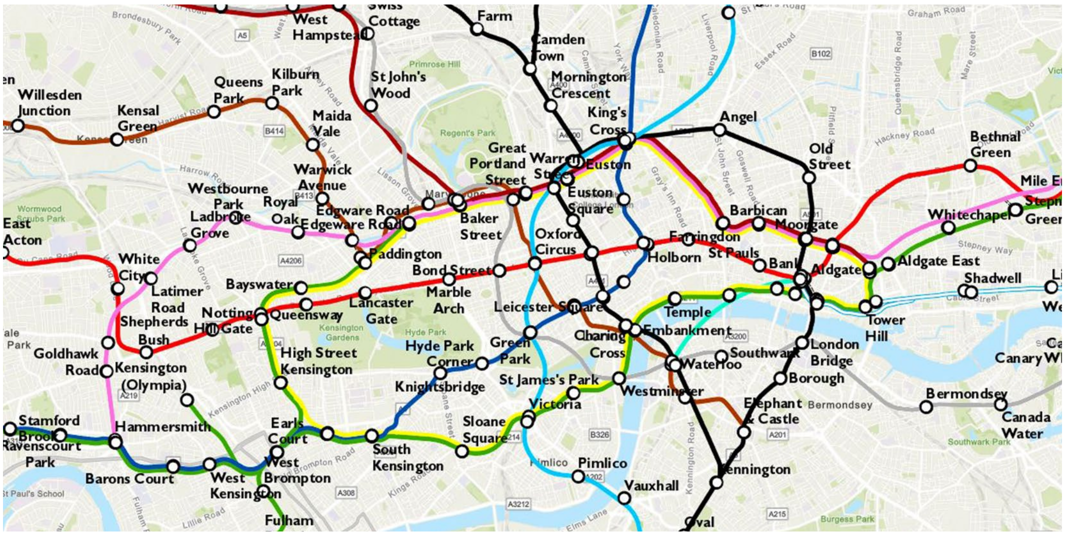


Figure 8.16. A geographically accurate map of the London Underground.

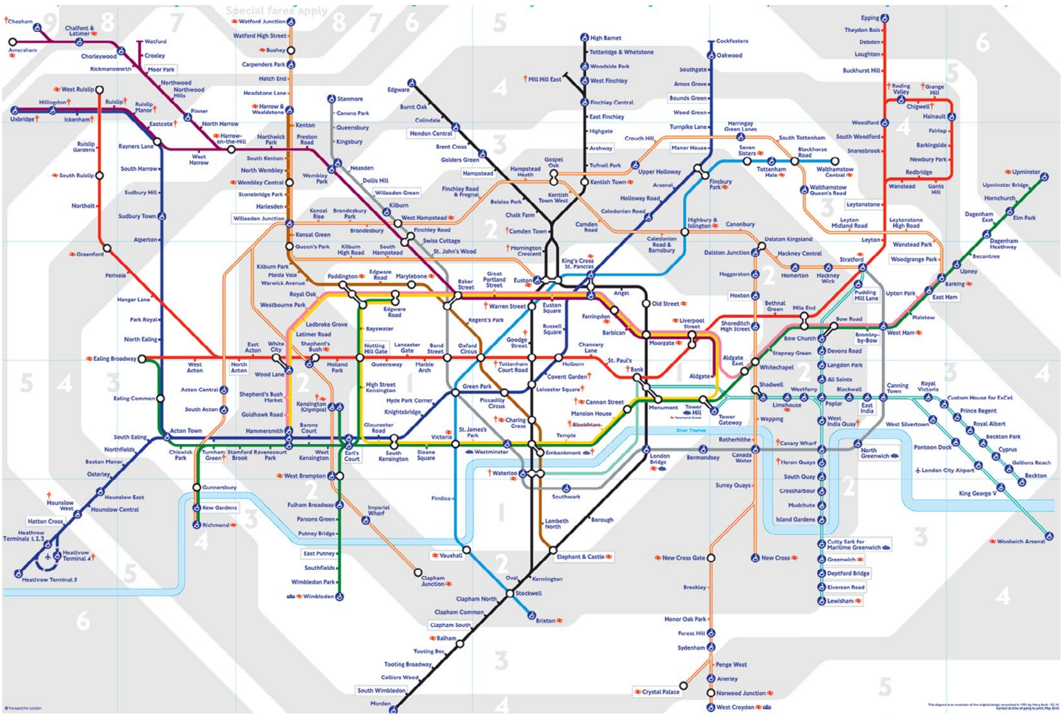


Figure 8.17 The map of the London Underground that is in use today.

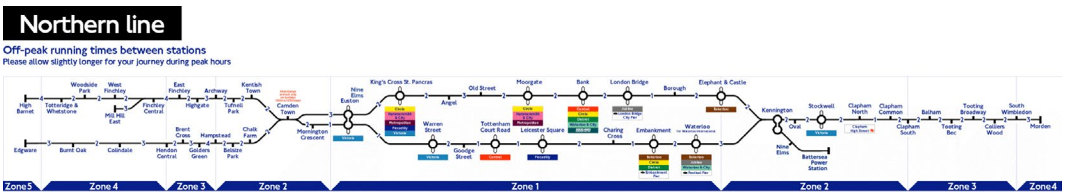


Figure 8.18 A straight-line diagram of a London Underground line as seen by passengers inside a train.

The above examples point to the value of condensing visual information, by eliminating non-task relevant features, but in a way that preserves or enhances clarity of meaning.

8.2.5 Personal and sociocultural aspects

Perception and interpretation of a sign partially depend on the user's prior experiences and on his socio-cultural background. Figures 8.19, 8.20 and 8.21 present three typical examples of the above effects.

Figure 8.19. The Rorschach Test: different people will interpret it differently depending on their proper life experiences.

It is well known from psychology that the image in Figure 8.19 has various interpretations depending on the person viewing it. Moreover, it has been observed that a person having seen this image and having interpreted it in some way, will seldom change this original interpretation if asked at a later time. It has also been shown that different interpretations of the particular image are not significantly correlated with the viewer's cultural background but rather with his prior experiences in life. Individual experience, in fact, forms a kind of cognitive substrate in interpreting the world, leading to diverse interpretations of the same sign by different persons.

Figure 8.20. An industrial product intentionally designed to suggest its target population (males / females) through its form and color.

Note that the geometry and functional characteristics of the two razors in Figure 8.20 are quasi-identical. Nevertheless, almost all respondents will identify the razor on the left with a men's razor and the one on the right with a women's razor. If asked why, respondents typically mention features such as the colors, the engraved patterns and the overall shape of the object, using expressions such as "more powerful", "stricter", "more angular" for the men's' razor and "more fluid", "more floral" and "more playful" for the women's one. Therefore, one can see a clear association of primitives such as strictness, power and rigidity with the masculine element, while associating flexibility, joy and nature with the female one. These associations of primitives are mostly socio-cultural in origin. They are "stereotypes", i.e., common beliefs embedded in the social consciousness of both men and women of a particular socio-cultural group. Stereotypes are thus intersubjective cognitive constructs that are impulsively employed by members of a socio-cultural group, without often the ability to offer a causal connection or explanation of such associations.

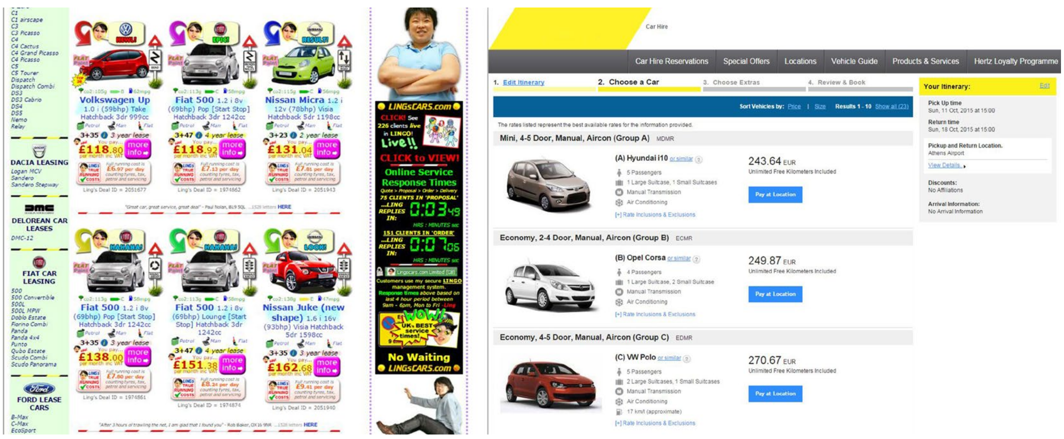


Figure 8.21. Internet homepages of two major car rental companies.

Figure 8.21 presents the homepages of two major car rental companies. The homepage on the left is addressed to customers in Southeast Asia, while the one on the right to customers of European origin. The reason why Europeans tend to trust the homepage on the right more than the one on the left has its origins on the cultural associations the Europeans have among the notions of strictness, clarity, reliability and, ultimately, trust. On the contrary, maximalism, diversity and visual intensity are not associated with precariousness or lack of quality, unreliability or mistrust in Southeastern Asian culture. It is easily seen here that the cultural–historical substrate of a person belonging to a particular socio-cultural group has engraved certain –often unacknowledged– connotations that are passed from generation to generation.

8.2.6 The role of context

Human meaning making is embedded in context, as much as it is embedded in individual history and sociocultural background. In other words, the same individual, having very specific prior experiences and socio-cultural substrate, will interpret a signal quite differently depending on the place and preoccupation (context) they find themselves at the very moment they receive the signal (see Figures 8.22, 8.23).

Figure 8.22. The utterance "close all the windows" takes a totally different meaning depending on the context.

Figure 8.23. This particular sign is found in the traffic code and signifies that "passage is prohibited". However, if seen on the can of an alcoholic beverage, its interpretation will change to a completely different meaning i.e., "don't drink while driving".

Context can be understood as having both objective and subjective elements. Objective elements are the immediate physical environment when one perceives and interprets a signal (e.g., an auditory signal in a noisy environment or a visual one placed outside the person's visual field), while subjective elements entail the individual's cognitive activity at the moment (e.g., the user preoccupied with his computer or with the ambient environment of a room). Therefore, when designing information, one should pay attention to how moment, place and preoccupation is possible to affect signification.

8.2.7 Paradoxes in signification

A sign is always implemented in a specific form. Thus, even small details of sign implementation may carry their own connotations, liable to interpretations of their own, in such a way that may blur, or even contradict, the intended sign interpretation itself.

Figure 8.24. Which is the signal and which the interpretant in the above picture?

For instance, in figure 8.24, at a first level one might correctly say that the signal is the array of letters painted on the wall, and the interpretant the prompt not to use the walls for writing slogans. At a second level of interpretation though, the implementation of the signal itself on a wall negates itself, thus will often be re-interpreted as irony. In theoretical terms, the first interpretant becomes the substrate of a second meaning making process. This cumulative signification process can go on in the interpreter's mind through what Charles Sanders Peirce has termed "thought signs" (Akin 2022).

The Investment Bank

The Investment Bank

Figure 8.25. Two alternative logos for the same bank.

In Figure 8.25 a bank sign is materialized with two different typefaces. Which one instils more trust to a prospective customer? The specific typeface employed in a logo will itself carry connotations to the receiver, beyond the semantics of the bank name it represents. In this specific example, the strict geometry and the serif style (the small features at the end of strokes within letters), of the top typeface is culturally connected with printed material of authority or prestige, while the second typeface mimics handwritten material and it is often used for comics or children's books. Therefore, in the second typeface, a contradiction arises between the sign content and its form. In practice, when implementing printed material, graphic designers will select a typeface family that is deemed culturally appropriate for the specific application.

8.2.8 Information presentation and cognitive load

The way information is presented has a profound effect on human cognitive processing. Consider the following two player number scrabble game:

Given the list of numbers 1, 2, 3, 4, 5, 6, 7, 8 and 9, each player is asked in turn to pick a number from the list. A particular number can only be picked once and is then removed from the list; each player may choose up to three numbers. Players may note down in writing the numbers they pick. The game is won by the first player holding any three numbers whose sum is 15.

The game is cognitively demanding since it requires from a player to try forming a triad adding to 15, while at the same time preventing her/his opponent from doing so before s/he does. To do so, players need to compute and anticipate all possible winning triads from the set of numbers they or their opponent possess or are about to choose. Eventually, after recurrent rounds a player may develop technics enabling her/him to outperform a novice, but still the gameplay rests cognitively demanding and subject to intermittent errors.

The number scrabble game can be visually transformed as shown in the Figure 8.26. Note that in the 3 x 3 board each horizontal, vertical, or diagonal triad adds-up to 15. Thus, in this version, the players, instead of computing and anticipating triads, may just choose triads conforming to a simple visual constraint (i.e., triads forming a straight line). The game may be further transformed by doing away with the digits altogether. It then becomes the well-known tic-tac toe (or noughts and crosses), a game that can be played by almost anybody, even preschool children.

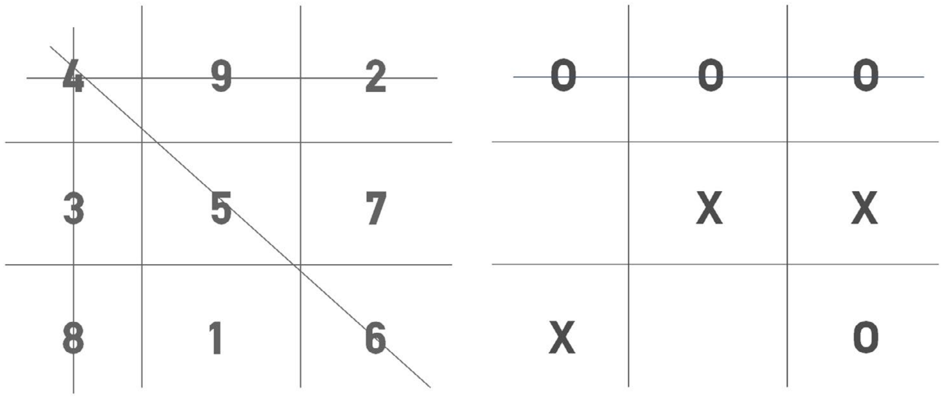


Figure 8.26. The number scrabble game transformed to the tic-tac toe game.

It is clear from above that the number scrabble and tic-tac toe games share the same underlying formal structure even though they are totally different in appearance. Games differing in appearance but sharing the same structure are called “isomorphic”. Why is it then that the tic-tac toe is considered by most adults as a rather trivial whereas the numbers scrabble cognitively demanding? Indeed, empirical evidence shows that playing tic-tack toe is faster; players take less time to make decisions and commit fewer errors. The answer lies in the game’s appearance; in the tic-tac toe, all needed calculations have been incorporated in the visual design of the game interface, doing away not only of the need for mental calculations but of recognition/remembrance of numbers altogether. Moreover, the visual interface of tic-tac toe enhances quick development of advanced tactics, resulting in novices attaining expert performance after just a few rounds.

One may thus suggest that the visual appearance of the tic-tac toe version of the game is an exemplary decision support interface for an otherwise demanding human cognitive task. Therefore, the way an isomorphic game is presented may profoundly affect gameplay performance depending on the cognitive abilities of the agent playing (human or machine). Note that for a digital computing machine it is far less resource demanding to play or solve the game in its numerical form. As a more general note, when designing the visual form of an object one should strive to adapt the design in such a way that makes the information more salient in terms of perception and interpretation for the target user and the task at hand. This typically involves hiding nonessential or redundant information, and presenting information in a format that needs the least cognitive processing in the user’s head.

The above principle applies in any type of visual representation. For instance, a weather forecast may be presented to a user in the form of a written text or through graphical means (Figure 8.27). The two representation forms are isomorphic in the sense that they describe the same phenomena through different means. However, the written text is perceived and interpreted sequentially. This sequential information reception inherent to reading requires retention in memory, and often makes the reader go through the whole text before arriving to a proper understanding of the weather forecast. In contrast, the isomorphic chart and weather symbol format is quicker to perceive and easier to interpret due to the nonlinear, pictorial form of information presentation. A person already acquainted with a particular geographic chart and relevant weather symbols may get an overall picture of the region forecast at a glance, and quickly focus on a particular area of interest without the need for scanning for place names. Note, also, that administrative area names do not fully correspond to map coordinates nor to particular weather patterns, making the chart format more adequate for conveying weather information over administrative segmentation.

```
Weather forecast for Friday 03.03.2023

The Ionian islands, eastern Crete and most
continental areas will be cloudy with local rain
showers. In the rest of the country, a few
clouds will increase gradually from west to the
east. A few snowflakes will fall in the northern
mountainous areas (above 1500 meters). The winds
will blow in the west from southern directions 3
to 4 Beaufort while in the south there will be
western winds 5 to 6 Beaufort. In the east,
winds will blow from northern directions 5 to 6
and in the north locally up to 4 Beaufort. The
temperature in the north will not exceed 12 to
14 degrees, in the rest of the country it will
reach 15 to 17 degrees and locally in the
Dodecanese, 19 degrees Celsius.
```

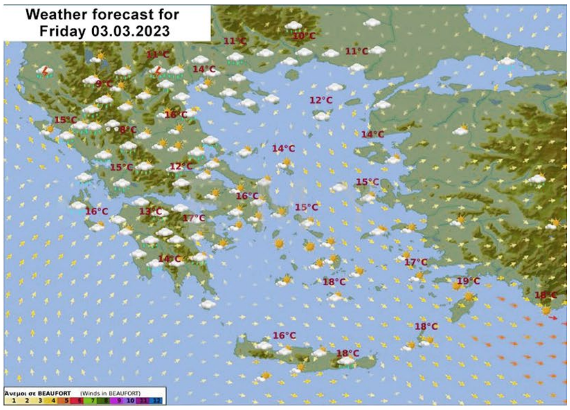


Figure 8.27. The same weather forecast in two different presentation forms .

8.2.9 Integrating visual design guidelines

The above guidelines, while presented independently, should, in fact, be considered in their complementarity since they display a high degree of interdependence. Note, also, that designing a visual interface is largely a creative process; therefore, as in any creative endeavor no single best way exists. Many alternative designs may satisfy the criteria set, leaving the designer to express aesthetic freedom. In any case, what is ultimately sought for is to convey the intended meaning in the most unambiguous form, by asking the following questions:

- What do we achieve to communicate?
- How many individual elements do we achieve to encompass in a coherent graphic configuration?
- How do these elements articulate? What level of detail or granularity is achieved?
- What blend between iconic / symbolic signs and cultural references does our design contain?
- Is there coherence between them throughout the solution?

8.3 Design principles for Controls and Displays

A contemporary trend in automobile dashboards for some years is to incorporate a touch-first digital interface to control the entire car's auxiliary systems such as the heating, sound, GPS and driving assistant features. Owners (and road testers), however, are starting to complain about laggy response, important functions buried in touchscreen menus, physical annoyances like capacitive sliders, not to mention visual distraction during driving. Although touchscreen interfaces have long been established in smartphones and notebooks, is seems that in automobile context physical controls outperform them in many respects.

Indeed, in November 2022, Thomas Schäfer, CEO of the Volkswagen Group, in an interview at the Los Angeles auto show, has admitted that the company's new touchscreen infotainment systems have not been good enough, and promised to roll out fixes which include the return of hard buttons at least for the 10 most used functions (Pollard 2022).

People interacting with machines (e.g., tools, computers, equipment) need to send and receive information to and from the machine. Therefore, the two main components of interaction from the human perspective are (i) acting upon, and (ii) receiving information from the machine. The elements we act upon machines are termed controls or controllers, whereas the elements we receive information from them are generally termed displays. Controls and displays are thus mediators linking human actions to system functioning.

Controls entail all information input means that human operators use to provide instructions to machines. Typical controllers are levers, keys, switches, keyboards, speech commands, gestures, etc. The display components of interface systems entail all means that machines use to provide information to human operators. Typical display means are visual displays, dials, lights, audio or vibrotactile signals, etc.

Design of controls and displays are closely linked, forming a human-machine interface. In some interface designs, controls and displays are clearly distinguishable, such as display dashboard and controls in an automobile. In others, controls may be embedded into display arrangements physically coupled, with the latter such as in Supervisory Control and Data Acquisition (SCADA) graphical interfaces or in touchscreen displays.

Displays and controls selection is heavily dependent on specific task requirements and human operator needs. In order to choose the most appropriate mix, the designer must first have a well-defined objective that can be summarized in the following questions:

- What information does the operator need to be displayed and why?
- What control actions are required by the operator and why?

Once these objectives have been clearly set, the designer needs to define the means of information display and system control. Various international or industry standards and guidelines exist for reference, depending on the domain of application (e.g., ISO 9355-2 Ergonomic requirements for the design of displays and control actuators or ISO/IEC 9995 Information technology - Keyboard layouts for text and office systems). Nevertheless, the following paragraphs provide some general insight.

8.3.1 Design of controls

Depending on the variable being controlled, controls may be discrete (two state, multiple state) or continuous (linear or progressive). For instance, keyboards are multiple state discrete controls whereas a computer mouse is a continuous non-linear control (because mouse speed has an exponential relation with cursor speed). In terms of input modality, controls can be motor/force or voice activated, or more exotic ones such as eye-gaze or recently electroencephalogram (EEG) activated (Figure 8.28).

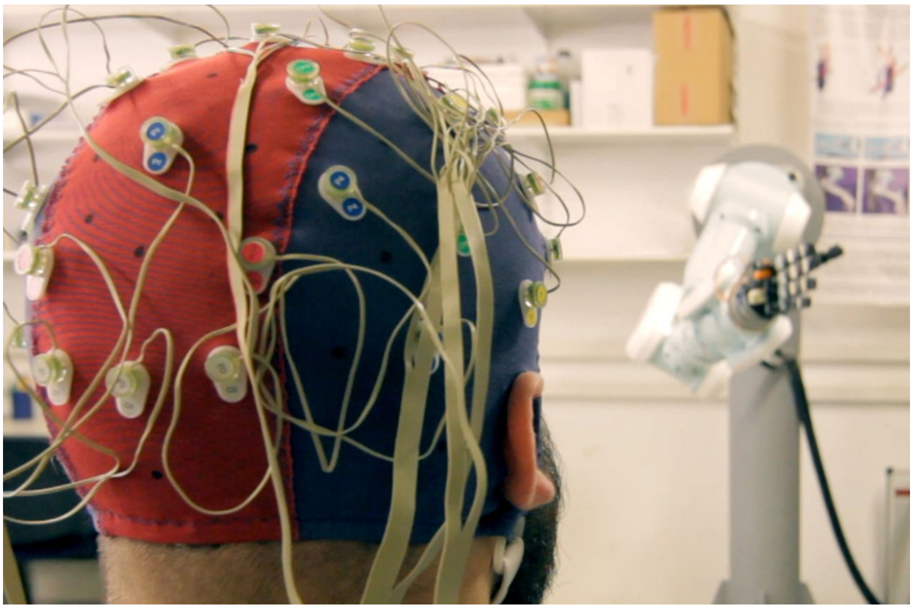


Figure 8.28. A researcher controlling a robotic arm through an EEG system in the Control Systems Lab of the National Technical University of Greece

Based on the specific task requirements and operator needs, a number of criteria should be considered before the final controller selection. An indicative list is provided below:

- Affordance of use, control range and status indication. The control's design must indicate the way it is meant to be used (rotate, slide, etc.), its operating range (i.e., where are the starting and final positions) and its current status (e.g., a two state ON/OFF toggle switch or a steering wheel provide visual feedback of their current status while a typical pushbutton does not).

- Feedback on status change (e.g., a physical button provides kinesthetic feedback by being recessed once pressed, while a virtual button may only provide visual or auditory feedback by color change or sound).
- Ease of detection of the control (e.g., a robot emergency stop button should be readily detectable –see Figure 8.29) and ease of discrimination among multiple controls (e.g., multiple valve controls in a chemical process interface should be easily distinguished from each other).
- Ease of operation of the control, subject to the required frequency and duration of use (e.g., the accelerator pedal in an automobile should be given priority in ease of use over the air-conditioning unit).
- The time criticality and required immediacy of control, (e.g., automobile brakes or aircraft eject levers should be designed to be actuated in the range of 100-200msec following the operator's intention to activate, whereas temperature setting for a water heater, or steering of a large ship do not require such immediacy since the system's response to control actions presents considerable latency).
- Prevention from accidental effectuation (e.g., force threshold activation of a gun trigger or touch duration activation threshold for a virtual touch button).
- The force required to be exerted should be related to the function it serves, (e.g., the steering wheel of an automobile should be force resistant to prevent accidental overturning to the point that it does not fatigue the operator).
- Whenever applicable, the incorporation of feedback should be considered in the controller to communicate the state of the system during its control (e.g., an airplane stick should provide force feedback of the ailerons resistance to communicate current aerodynamic pressure on them).
- Natural mapping, (i.e., compatibility with the resulting effectuation to the controlled system), is desirable where applicable, (e.g., turning a steering wheel leftwards should displace the vehicle towards the left). Note that some traditional controls, e.g., boat tillers, do not adhere to this criterion and thus are often challenging to operate by novice users.
- Compatibility with the required precision of the controlled system (e.g., a sound volume knob requires a resolution of 1dB since higher resolutions are marginally intelligible to the human ear; the same can be said for room temperature control where less than 0.5° C difference is not intelligible to the human).

- Proper design respecting the context of use (e.g., potential use of gloves or other gear might decrease the user's dexterity or feedback perception). Moreover, the need for control in situations of high acceleration forces, or vibrations should not affect the user's sense of control.
- Customization options, such as freedom of assigning functions to controls and sensitivity response, when critical for expert performance.

Figure 8.29. A manual robot controller teach pendant. The emergency stop button is readily detectable through its position, size and color. What is more, to prevent its accidental effectuation a rotating action is required instead of a pushing action.

Take the example of comparing a physical knob versus a touch screen button to control the cabin temperature of a car. The given scenario is to increase the air temperature by two degrees while driving alone on a slightly congested road. Considering the task, we can assume that the manipulation should be done one handed and with minimum visual distraction from the road. The context of use also implies that there might be vibrations and accelerating forces while manipulating the control. Consider the two alternative controls shown in Figure 8.30. The first is a rotary knob with small grooves which increases the temperature when rotated clockwise and decreases it when rotated counterclockwise. The knob has haptic steps, each changing the temperature by 0.5° C, while the selected temperature is being shown on an LCD above the control. The second control is embedded on the bottom left corner of a multimedia touchscreen, consisting of a plus and a minus sign which change the temperature accordingly by tapping on them, with also a step of 0.5° C. Therefore, the first must be rotated clockwise by four steps while the second must be tapped four times on the plus sign.

Figure 8.30. Two alternative controls for cabin temperature in a car, a physical knob (left) and touchscreen buttons (right)

In terms of ease of detection, the rotating knob is far easier to spot, not only due to its bigger size but also due to its tangibility (i.e., its form stands out), which has the advantage of being located haptically without the need of visual contact. The touch control on the other hand, requires visual contact to locate, and needs more motor accuracy to manipulate due to its smaller size and lack of tactile feedback. What is more, since the action must be done inside a bumpy cabin, the tapping on such a small target might cause accidental effectuation of the adjacent controls, while the grabbing of a physical knob can help withstand such environmental disturbances. Moreover, the potential wearing of gloves certainly hinders the effective use of the virtual button.

Regarding their affordance of use, the rotary knob certainly communicates its rotating function through its shape, while the touch button is less conspicuous in terms of possible manipulation. Both controls lack indication on their control range while they do inform on their current status indication only through their accompanying display.

The change of heating status while acting upon the control cannot be instantly perceived, due to the slow response of heating, unlike, for instance, a change in the volume of a sound system. Therefore, haptic feedback on this control is useful, so that the driver can perceive an input response from the system without taking his eyes off the road. The rotary knob, through its haptic steps, satisfies this requirement but not the touch button, which only provides visual feedback.

While the physical knob is superior in terms of usability, on the other hand, the virtual control offers greater versatility in terms of design, since it can be updated in size, position and accuracy giving room for customization. The virtual control, also, occupies less space than the physical knob, leaving room for more functions on the whole.

8.3.2 Display design

Depending on the nature of information to be conveyed and the task at hand, displays may present discrete (two state, multiple state) or continuous data, but also configured multi-dimensional data (e.g., 3D or plotted variables) or semantic (e.g., text). In terms of sensory modality, the most commonplace information displays are visual, however auditory, haptic/vibro-tactile or even olfactory displays can also be used depending on the application (Figure 8.31). Again, the specific cues to be employed in the design of displays depend heavily on the task requirements and the needs of the human operator.

Figure 8.31. In cave diving where lighting is exclusively depended on artificial sources, a mere visual wayfinding system or map would be pointless and even dangerous. Instead, the use of physical strings and haptic markers of several shapes are used for orientation as both visual and tactile references. For example, arrow markers – as the one shown here - point along the line leading to the nearest surface with breathable air (Photo on the left by SJ Alice Bennett).

An indicative list of criteria for display type design/selection is provided below:

- Information should be in a format that is appropriate for the cognitive activity it is meant to support. For instance, the qualitative level indicator (pictorial) display for a tank level (Figure 8.32 center) will be more efficient for gross checking than a purely alphanumerical one (Figure 8.32 left), as it visually presents level difference from min and max values without the need for mental calculation. A dynamic indicator over time can even show trend and relieve the user from remembering past values (Figure 8.32 right). On the other hand, a quantitative alphanumeric display may be more appropriate for automobile speedometers, since absolute value reading allows direct comparison of current speed to the road speed limit, while a dynamic analog[9] indicator may be more appropriate for automobile tachometers (i.e., rpm gauge) for gross monitoring of engine rpm operational range (Figure 8.33).

- Information should be easy to perceive and should prevent from erroneous interpretation. For example, a sound alarm is a good choice for conveying critical information, but it should be designed to be easily detected in a noisy environment and be correctly interpreted as such. Moreover, a combined auditory and vibro-tactile signal is a good choice for emergency human take-over signal of an autonomous car, as these sensory modalities need less conscious observance and do not interfere with any other visual task of the driver.
- The design should assist the operator to create a mental model as close as possible to the actual system (e.g., Figure 8.34, Figure 8.35).
- Information value scales should have no more range and precision than required for the task. For example, the set-up of digital wake-up alarm should provide no more than 1 min precision, since it is seldom the case that one needs second level accuracy for their wake-up notice. Providing excessive precision (e.g., an alarm clock set at 07:05:09) may even cause erroneous set-ups.
- Changes in the information status of displays should be easily detectable.
- Display malfunctions should be readily detectable.

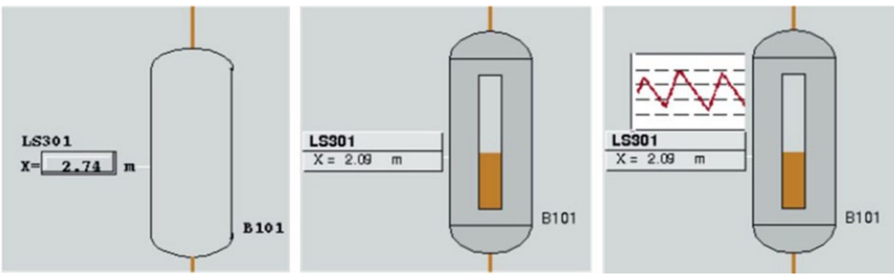


Figure 8.32. Display with (a) digital (b) digital / analogue and (c) digital / analog / dynamic (time-trend) tank level feedback. Source: University of Oldenburg.

[9] The terms "analog" and "digital" used here do not refer to the technology connecting the system with the display (i.e., electronic or mechanical wiring), but it is rather to the difference between alphanumerical instantaneous data and dynamic level indicators.

Figure 8.33. A digital alphanumeric speedometer, next to an analog rpm gauge. Alphanumeric values are best suited for observing speed limits while analog readings are best suited for rpm range monitoring.

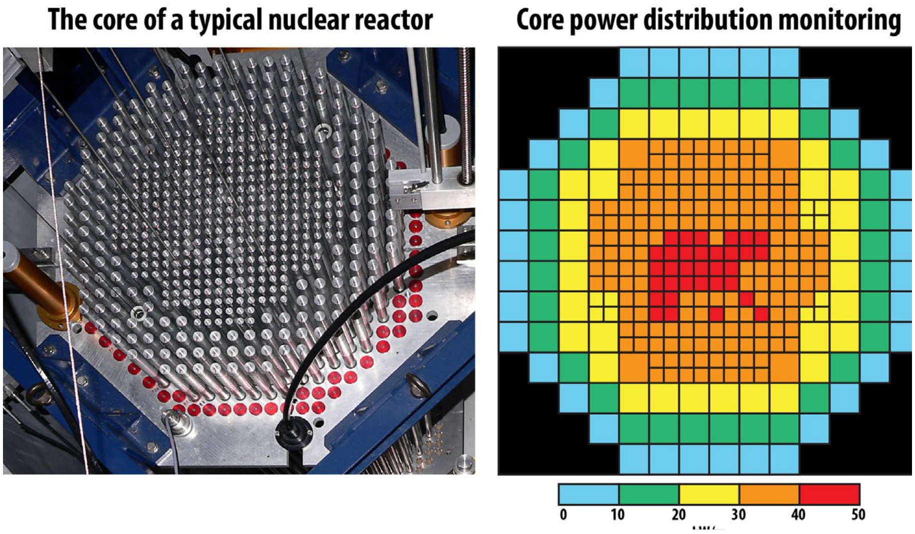


Figure 8.34. On the left, the rods' array of a typical nuclear reactor and on the right a potential monitoring system of the distribution of the core's power.

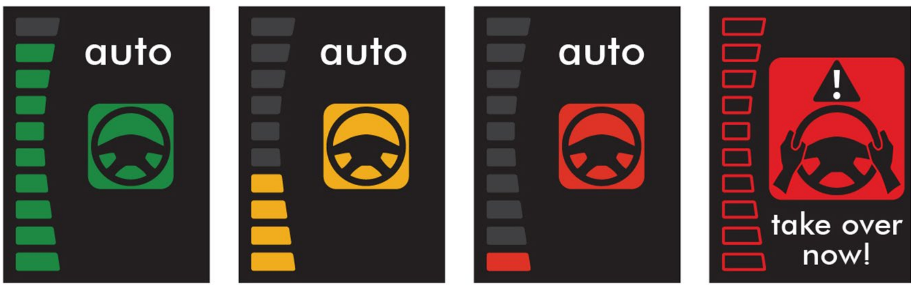


Figure 8.35. An experimental interface for autonomous cars, developed in the National Technical University of Athens. The colored bar shows the level of "confidence" of the car automation to respond to environmental conditions, thus assisting the driver to correlate the road conditions with the internal driving algorithms of the car and therefore, create a more coherent mental model of its function.

The above criteria for control and display selection are only indicative. Depending on the specifics of each design problem more specific criteria may also apply. Finally, it should be noted that, as in any design effort, it may be impossible to equally satisfy all criteria. Therefore, in any particular application one should make judicious trade-offs.

8.3.3 Configural displays

Human operator performance can be improved by providing displays that allow the user to exploit the more efficient processes of perception and pattern recognition, instead of requiring them to utilize the cognitively intensive processes of memory, calculation, and inference. Configural displays (also called Ecological displays) are typically used for integrating many discrete information sources in one general picture. As a case in point, Woods, Wise, and Hanes (1981) developed a display that presents information concerning the general operational status of a nuclear power plant where more than 100 individual sensor values are mapped into an octagon (see figure 8.36). Mapping multiple process variables into a single geometric form creates higher level visual properties, such as closure and symmetry. Such configural displays are currently in use in several nuclear power plants.

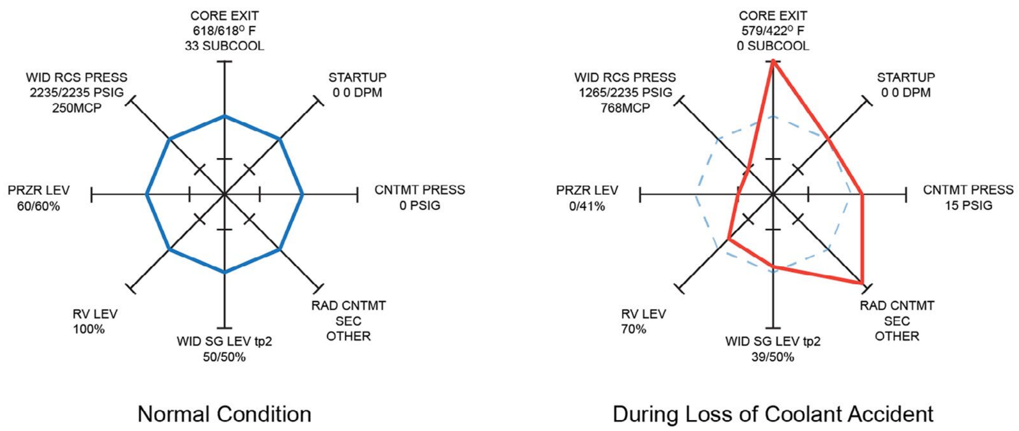


Figure 8.36. Configural display presenting information on the operational status of a nuclear power plant in two states of the plant (adapted from Woods et al. 1981).

The "building blocks" of a configural display are individual measured variables (that represent low-level data) in a domain. These low-level data are combined and carefully arranged in space so that they configure higher-order geometrical forms. Variations in the values of the measured domain variables will, therefore, produce distortions in the shapes of the higher-order geometrical forms. These distortions are usually referred to as "emergent features" or higher-level visual properties (e.g., symmetry, parallelism) that arise from the interactions of the lower-level graphical elements.

The success of such a design depends on identifying the semantic properties of the work domain (e.g., mass – energy equilibrium, combined constraint limits) and on mapping their functional dependencies into geometrical configurations.

A properly designed configural display will allow the trained operator to see the nature of a problem directly, using powerful perceptual processes as opposed to deducing a diagnosis, using the less efficient cognitive processes, such as mental comparison and several calculations.

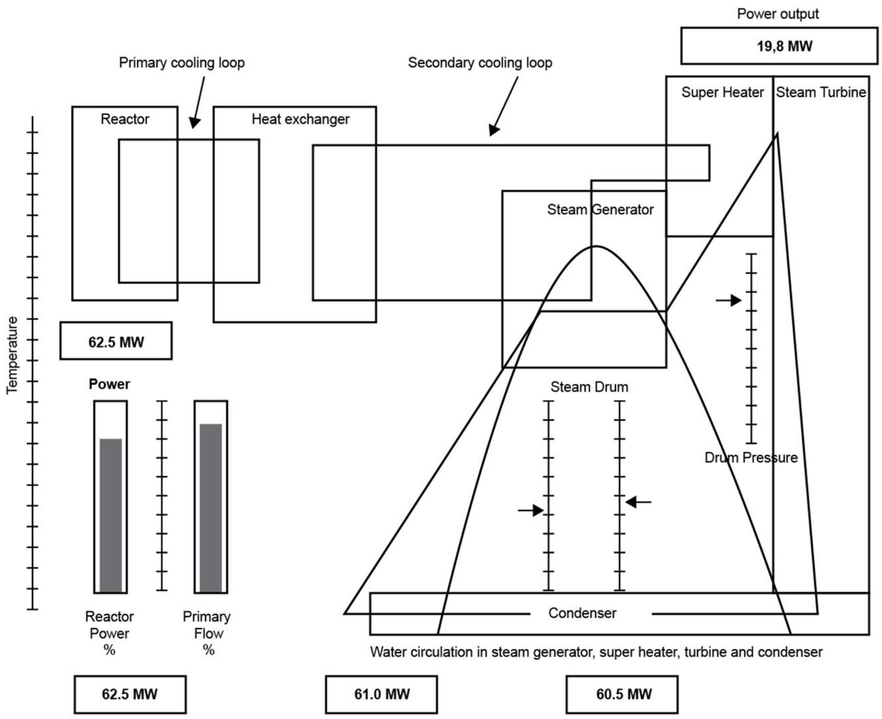


Figure 8.37. An ecological interface for monitoring the operation of a nuclear power plant. (adapted from Lindsay and Staffon, 1988).

Figure 8.37 shows another early configural interface destined for nuclear plant control designed by Lindsay & Staffon (1988). The display presents a real-time thermodynamic model of the plant processes providing a direct indication of plant performance to the trained eye. Individual variables, e.g., temperature levels of primary and secondary coolant loops, as well as inlet values for steam generator, super-heater, turbine and condenser are provided separately in typical bar form. However, the overall plant status can be directly perceived as an emergent feature through the graphical arrangement of the individual variable graphs. The Rankine cycle (on the right side of the display) presents inlet constraints graphically, in a way that helps operators directly perceive reactor status, based on visual patterns rather than infer reactor status based on cognitive information processing of individual variables. Sudden changes in the installation or sensor faults are readily noticeable as deformations of the rectangles representing the primary and secondary cooling loops (top left) or as temperature values exceeding the Rankine cycle boundaries. It is important to stress that complex configural displays may not be readily grasped by the lay person; they are destined for professional operators and may require considerable training before becoming effective in use.

8.3.4 Configuration of controls and displays

Multiple displays and controls are usually combined to create an interface. The term interface, here, is understood in its broader sense of the overall configuration, which connects a user with a system or machine, and not exclusively on graphical interactive displays. In this sense, an industrial crane cockpit, a sensory glove and googles for microsurgery or a ship engine control-room (Figure 8.38) are all configurations of displays and controls that form interfaces. The proper configuration of displays and controls is a critical factor for human – machine system effectiveness.

Figure 8.38. View of a ship engine control room: a walk-in interface between man and machine.

Various design principles exist for tackling interface design, such as functional groupings or task sequence arrangements of controls and/ or displays. However, these tend to be highly specific to the application domain. Some general criteria for the relative arrangement of controls and displays in an interface are provided below:

- Important and/or the frequently used displays/controls should be placed centrally (or on the home-screen if applicable).
- Information displays often used together should be grouped consistently with the sequence in which they are typically consulted (arranged from left to right and/or top to bottom according to the checking sequence).
- Information displays should be linked to the corresponding control functions.
- There should be a topological correspondence between display/control placement and the arrangement of the physical processes that these represent.
- Mimic diagrams (see Figure 7.6) should be used when the topology of a process is relevant for task execution. Mimic diagrams have a pervasive effect on operators' mental models of a process, thus special attention should be given in their design.

Though the above criteria provide some direction, the complexity of the subject makes it that interface design is highly specific. Depending on the domain of application, there are various standards and/or guidelines, e.g., ISO 11064, Ergonomic design of control centers, graphical interface guidelines for Electronic Chart Display and Information Systems (ECDIS) on ships, or user interface guidelines for various medical devices. However, despite standardization for many domains, control and display design still entails a predominantly empirical flavor (i.e., through iterative testing) and is partially an applied art.

References

Atkin, A. (Fall 2022 Edition) «Peirce's Theory of Signs», *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/fall2022/entries/peirce-semiotics/

Lindsay, R W, and Staffon, J D. *Model-based display system for experimental breeder reactor II*. United States: N. p., 1988. Web.

Naser, J., Fink, R., Hill, D., & O'Hara, J. (2004). *Human Factors Guidance for Control Room and Digital Human-System Interface Design and Modification*. EPRI TR-1008122. Palo Alto, CA: Electrical Power Research Institute.

Pollard, T., (2022, November 19). *VW shakes up infotainment touchscreen UX to fix mistakes of the past.* carmagazine.co.uk. Retrieved February 13, 2023, from https://www.carmagazine.co.uk/car-news/tech/vw-infotainment-and-touchscreen-problems/

Woods, D. D., Wise, J., and Hanes, L. (1981). An evaluation of nuclear power plant safety parameter display systems. *Proceedings of the Human Factors Society 25th Annual Meeting (pp. 110-1 14)*. Santa Monica, CA: Hu- man Factors Society.

Chapter 09: MEDICO - a User Centered Design process case study

Chapter 09: MEDICO - a User Centered Design process case study

Chapter Summary

This final chapter aims to integrate the various methods presented in the book through a case study of User Centered Design process in a real-world project. Specifically, it presents the development of an innovative emergency medicine IT application starting from needfinding and negotiations with stakeholders for specification of requirements through iterative prototyping, information design, testing and implementation. The aim is to help the reader consolidate the knowledge acquired in the preceding chapters in an integral manner providing context and continuity to the successive stages of UCD.

Prerequisite knowledge

Basic knowledge of engineering concepts and the comprehension of previous chapters.

9.1 The Field

Figure 9.1. Providing remote medical advice to merchant ship crew, requires critical decision making that might lead to ships rerouting or emergency transits by air rescue.

The present case-study demonstrates the various phases of the UCD process involving the design of a collaborative information technology system (Nathanael & Marmaras 2000). The aim of the system in question was to assist the management of medical tele-consultations, provided by a Maritime Medical Advice Center (MMAC). The specific project was part of a European Research & Development program on Telemedicine and was conducted by one of the authors of this book in the late 90s. Although quite old, it is a typical example of a complete UCD endeavor, presenting the process in an integrated manner from initial user needs analysis to users and system requirements specification, prototype design and testing, and finally, to detailed design of the resulting solution given. The design team included two ergonomists, two software developers and two members of the user group.

Some background on the MMAC organizational environment at the time of the design process is the following. First and foremost, the IT application had to be introduced in a complex and sensitive work environment: the medical tele-consultation. Secondly, the prospective users were not enthusiastic in using information technology. In fact, a previous system was abandoned some years ago due to the unwillingness of the MMAC personnel to use it, apparently, because of its poor usability.

At the time of the design intervention, MMAC was the official Greek body offering remote medical advice to Greek speaking seafarers in vessels sailing worldwide. MMAC had been providing medical advice to vessels at sea since 1962. These services, however, were at first provided unofficially, through public medical emergency lines as this was convenient for the operations center of the Greek Merchant Marine Service. MMAC, as a dedicated Maritime Medical Advice Centre, officially started its operation in 1987 and gained legal status as the official Greek Medical Advice Centre in 1994. The service was free of charge for all vessels under Greek Flag. MMAC operated from the city of Athens and communicated with vessels mainly by vocal link through marine radio or Inmarsat.

9.2 Need-finding process and results

The methodology for the design of the IT system followed a user centered approach. At a **first phase** a need-finding process was carried out, conducted by two ergonomists, through an exhaustive work analysis of the existing situation. The techniques used for needs finding were a series of direct field observations averaging up to 60 hours including the shadowing of 30 medical calls, interviews with all the prospective users and an analysis of 500 old Medical Call Forms records.

9.2.1 The main Services provided by MMAC

The operation center of MMAC services concerning vessels at sea can be summarized as follows:

- Provision of direct medical advice to vessels at sea.
- Training and information services to mariners.

- Co-operation with, and advice to rescue authorities.
- Provision of medical certificates for the cases treated.

9.2.2 Detailed Task Description

A Hierarchical Task Description of MMAC is presented in Figure 9.2.

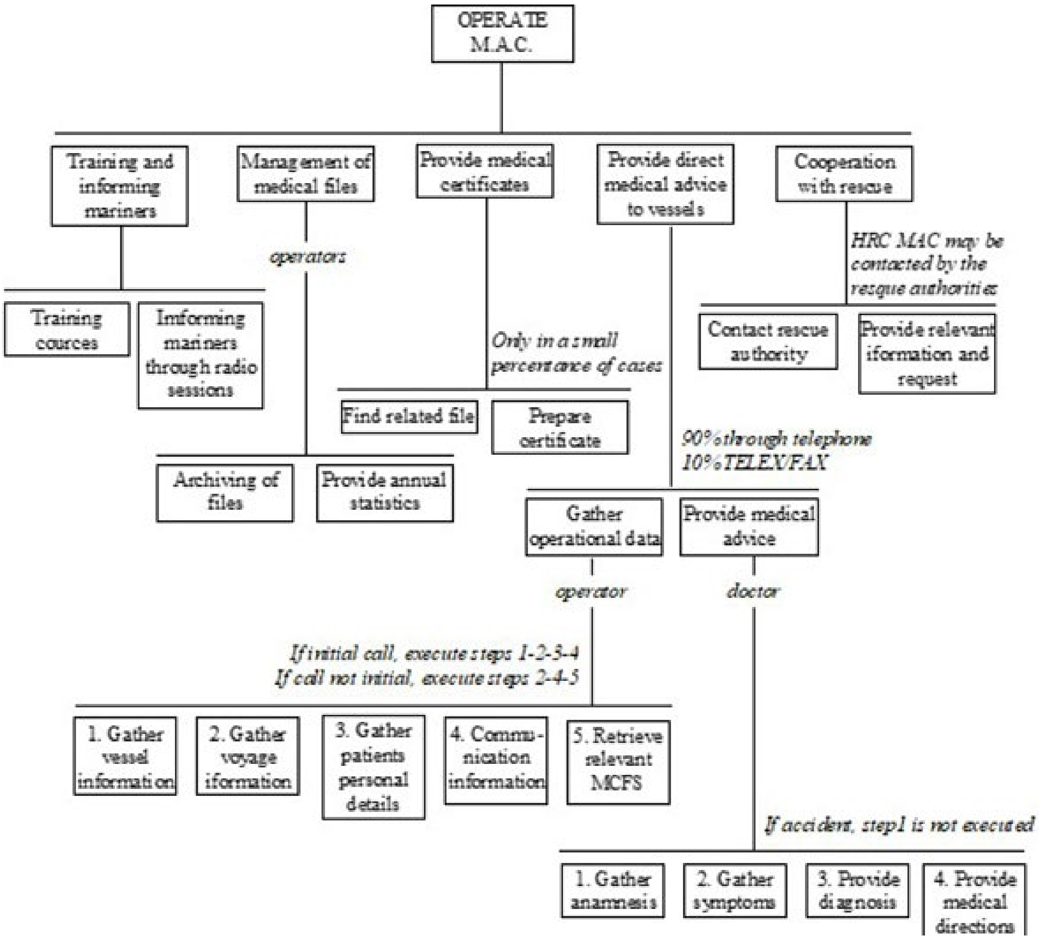


Figure 9.2. MMAC overall Hierarchical Task Diagram.

The information management system to be designed was intended to support three of the five main subtasks of the center's activity, namely; Provision of direct medical advice, Provision of medical certificates and Management of medical files.

9.2.3 Categories of vessels seeking medical advice

A rough estimation of vessel type seeking medical advice from MMAC during its 10 years of operation, were the following:

Merchant ships	80%
Fishing ships	10%
Liners & cruise ships	10%
Leisure craft	0%

9.2.4 Cases handled per year

From 1987 and up until 1996, MMAC had carried out 10379 radio-consultations, of which 83% concerned Greek speaking mariners. The yearly number of medical consultations has been constantly increasing since 1987 by an average of 18%. In 1996, 1852 radio-consultations had been carried out.

9.2.5 The personnel

At the time the new IT system was requested by the Center, MMAC personnel consisted of six telecommunications operators (TOs) and seven Medical Doctors (MDs). One TO and one MD were on duty in the operations center on 24-hour basis. All seven MDs were experienced in medical tele-consultation, having also served in urban medical call centers before joining MMAC. In terms of IT skills, five out of seven had no computer experience whatsoever. TOs professional competencies included computer skills (word processing and data entry tasks) and the operation of telecommunication devices (Telex, Fax, Radio etc.). TOs had no paramedic training. All personnel, MDs and TOs, received an initial training of 600 hours over 3 years in maritime matters. The quasi-totality of the personnel remains unchanged from the year of service uptake.

9.2.6 The medical consultation process

MMAC operated in a similar way as other medical emergency call centers, assessing the gravity of a medical situation through voice connection, and offering remote diagnosis and medical treatment along with deciding on proper evacuation or other medical aid.

However, since most of the cases handled by MMAC involved patients in the open sea, far from any land-based medical structure, its task included a more detailed assessment of the medical situation and provided possible diagnosis, medical directions for treatment and monitoring of the patients' health condition, until the ship reached a port with proper facilities. As a result, unlike a typical medical emergency call center, a medical case in MMAC usually involved several tele-consultation sessions, lasted for several days, and thus was usually handled by more than one Medical Doctors.

TO and MD sat in the same physical space, one opposite to the other and both communicated directly with ships (Figure 9.3). Typically, the TO intervened first in a call for medical assistance, and was responsible for gathering administrative information. Information was recorded directly in a hard copy of a 'Medical Call Form' (MCF) by hand. The TO, after recording administrative information on the MCF, passed the telephone line along and the MCF to the MD. In cases where the TO diagnosed (from acquired experience) a time critical emergency or high level of anxiety from the calling vessel, he could immediately pass the call to the MD, and ask for administrative information at the end of the communication.

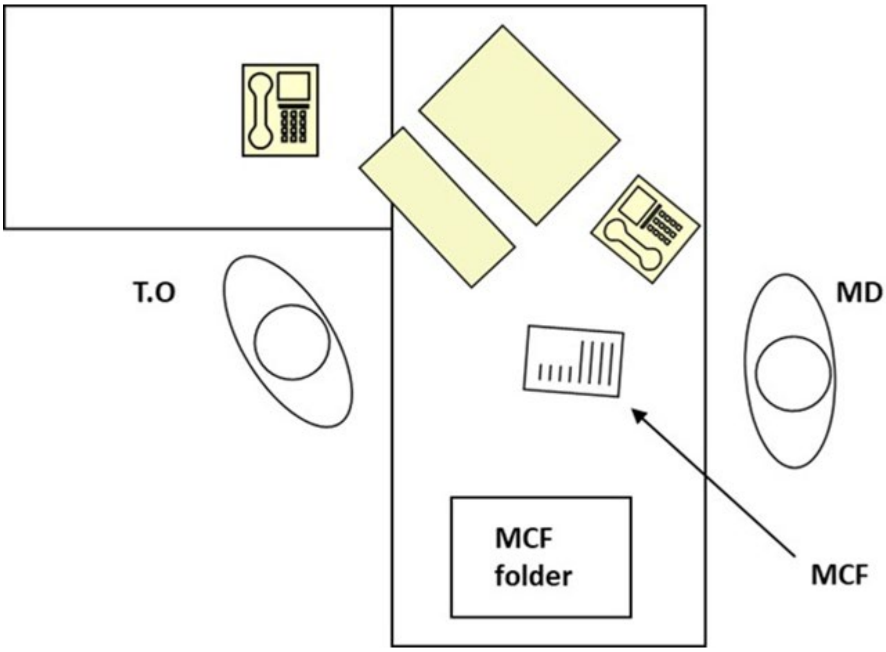


Figure 9.3. Physical arrangement of the MMAC workstations. MD (Medical Doctor), T.O. (Telecommunications Operator), MCF (Medical Call Form).

The vessel interlocutor was usually an officer and not the patient himself. If possible, and depending on specific conditions, the MD could ask to bring the patient next to the officer or to speak to the patient directly, but this was infrequent in practice.

In order to effectively undertake the medical consultation, the MD needed to consult in parallel, both the current and possibly the past MCFs. To find past MCFs for the same medical case and pass them over to the MD, the TO had to search the MCFs archive. However, for various reasons, MCFs were archived according to the time of call and not according to the case. This meant that retrieval of all previous MDFs for a specific medical case was laborious and not reliable. In practice, TOs were rarely searching through the archive, usually relying on the "vessels memory" in order to locate pervious calls (e.g., by asking the vessel interlocutor if he had communicated with them before for the specific case and when).

The MD typically proceeded by asking specific questions on the medical situation, the symptoms, the patient's medical history, etc. The process was not linear and it continued until the MD formed a 'satisfactory' knowledge of the patient's medical condition. On the basis of this investigation the MD provided the vessel with medical directions that consisted of treatment on board, medication, re-contact request or patient evacuation request. During the medical consultation the MD would write down in the MCF medical details on the case. This recording, if not comprehensive, could be completed after the consultation was over.

9.2.7 Monitoring therapy on board

In many cases the MD needed to have frequent updates on the patient's medical status. These updates followed the medical consultation course, and the results were recorded in a new MCF as described above. The contact was almost always initiated by the vessel; MMAC was financially restrained and could not call a vessel for medical updates. Only in a small minority of cases, where the concerned vessel was in Greek waters, MMAC would contact the vessel in its own initiative, via Athens' radio. Thus, there was often a concern about an evolving case, as MDs having previously managed a case had to go through the archived MCF files, one by one, to check if a vessel had subsequently contacted the center for an update.

9.2.8 Medical evacuation requests

In cases where the MD advised to evacuate a patient, MMAC could proceed either in a request for vessel rerouting towards the nearest convenient port or in issuing an official medical evacuation request to the Greek or foreign authorities depending on the vessels' position. This request could be done by telephone, telex or fax, upon convenience. Under Greek law, MMAC was the only national medical body with the formal authority to request medical evacuations from respective Search & Rescue authorities worldwide.

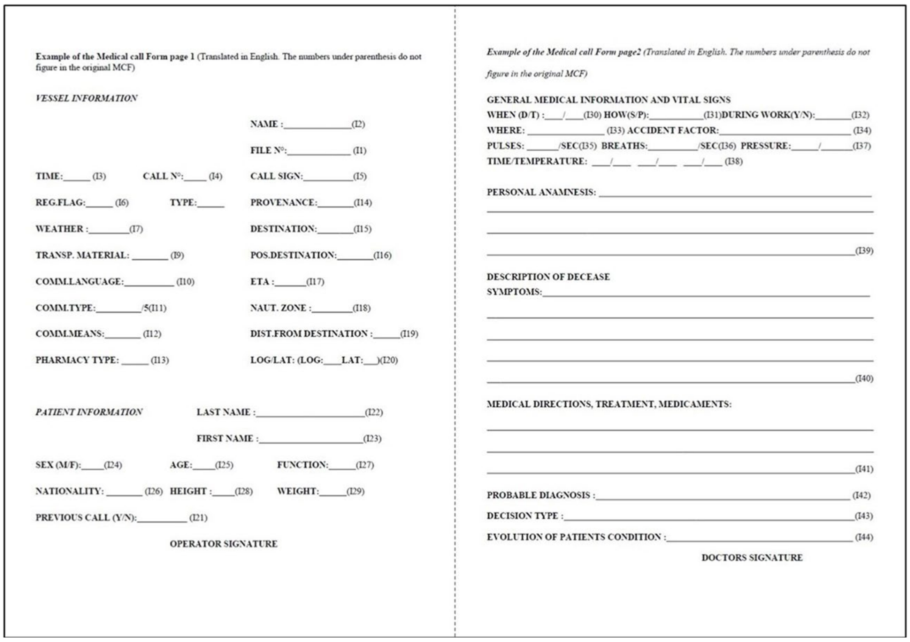
9.2.9 Provision of illness and accident certificates

The MMAC was authorized to provide accidents or illness certificates for the cases treated, upon the patient's request. These certificates had official status, and were prepared by an MD familiar with the specific case. Information for the certificates was retrieved from the archived MCFs. In such cases MMAC personnel had to go through the archive, find all related MCFs, "remember the case" and come to a medical conclusion, even a year after the specific event.

9.2.10 The Medical Call Form (MCF)

The MCF was a standard A4 format sheet of paper with predefined fields on both sides; one side for administrative and one side for medical information (Figure 9.4). The administrative part was composed of 29 (thus gathered by the TO) and 15 fields considered as medical information (gathered by the MD); a set of 44 information fields in total. This information set was the outcome of successive alterations by MMAC from hands-on experience after an initial set, which had been adapted from an Urban Medical call center.

An Extract from the observation notes from the UCD need-finding phase can be found in ANNEX I.



Example of the Medical call Form page 1 (Translated in English. The numbers under parenthesis do not figure in the original MCF)

VESSEL INFORMATION

NAME :____________(I2)

FILE N°:____________ (I1)

TIME:_____ (I3) CALL N°:_____ (I4) CALL SIGN:__________(I5)

REG.FLAG:_____ (I6) TYPE:_____ PROVENANCE:_______(I14)

WEATHER :________(I7) DESTINATION:_______(I15)

TRANSP. MATERIAL: _______ (I9) POS.DESTINATION:_______(I16)

COMM.LANGUAGE:__________ (I10) ETA :______(I17)

COMM.TYPE:_________/5(I11) NAUT. ZONE :________(I18)

COMM.MEANS:_______ (I12) DIST.FROM DESTINATION :_____(I19)

PHARMACY TYPE: _____ (I13) LOG/LAT: (LOG:___LAT:___)(I20)

PATIENT INFORMATION LAST NAME :____________________(I22)

FIRST NAME :____________________(I23)

SEX (M/F):_____(I24) AGE:_____(I25) FUNCTION:_____(I27)

NATIONALITY: _______ (I26) HEIGHT :_____(I28) WEIGHT:_____(I29)

PREVIOUS CALL (Y/N):__________ (I21)

OPERATOR SIGNATURE

Example of the Medical call Form page2 (Translated in English. The numbers under parenthesis do not figure in the original MCF)

GENERAL MEDICAL INFORMATION AND VITAL SIGNS

WHEN (D/T) :___/___(I30) HOW(S/P):__________(I31)DURING WORK(Y/N):_______(I32)

WHERE: _______________ (I33) ACCIDENT FACTOR:_________________________ (I34)

PULSES: ______/SEC(I35) BREATHS:_________/SEC(I36) PRESSURE:_____/______(I37)

TIME/TEMPERATURE: ___/___ ___/___ ___/___ (I38)

PERSONAL ANAMNESIS: ______________________________ (I39)

DESCRIPTION OF DECEASE

SYMPTOMS:______________________________ (I40)

MEDICAL DIRECTIONS, TREATMENT, MEDICAMENTS:

______________________________ (I41)

PROBABLE DIAGNOSIS :______________________________ (I42)

DECISION TYPE :______________________________ (I43)

EVOLUTION OF PATIENTS CONDITION :______________________________ (I44)

DOCTORS SIGNATURE

Figure 9.4. The paper Medical Call Form (MCF) used for recording medical calls in MMAC (Left - Administrative information, Right - Medical information).

To assess (i) the appropriateness of the MCF for the teleconsultation process, and (ii) the quality of information recorded in them, a detailed analysis of 500 past MCF records was performed. This analysis revealed a number of problems associated with their use. Specifically, on the administrative information side, of the 29 information fields, six (6) fields were never used in practice, while another four (4) seemed overlapping and ambiguous, since different operators would use them in different ways. For instance, the filed "Estimated Time of Arrival" of vessel to destination port was mostly recorded on time units (e.g., in two days) instead of the appropriate date units (e.g., 3/12 at 18:00 UTC); also, vessel position in geographical coordinates (Longitude/Latitude) was almost never filled in. The analysis, also, showed that no other information was kept in an informal way in the administrative side of the MCFs, i.e., side notes were rare.

On the medical information side, of the 15 information fields, only 5 – 6 were actually used in practice, the rest were almost always left blank. Table 9.1 summarizes the findings on the medical side of the MCF.

Table 9.1. The actual use of medical fields in the MCF

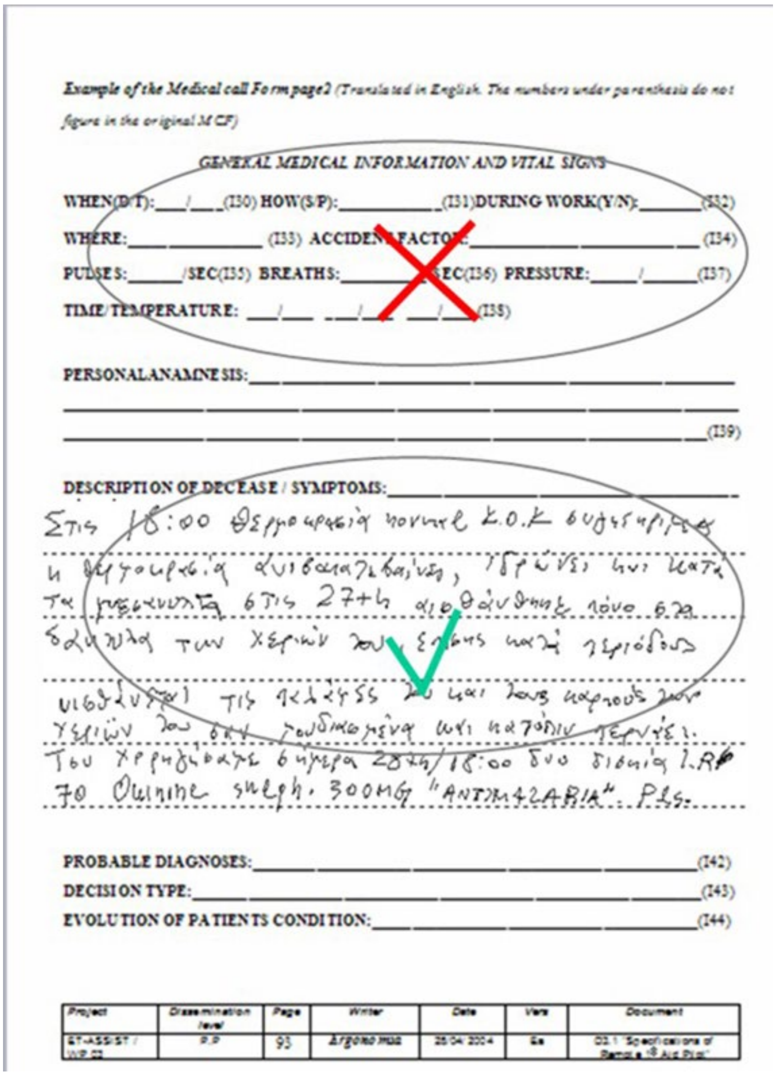
ID	Information field	Type	Description	Actual Practice
I30	When troubles started	Text	The temporal span of the medical troubles (the patient's estimation)	Never filled in.
I31	How troubles started	The patient's perception on the aggravation of the medical troubles	Never filled in.	
I32	During work	Y/N	If the troubles started during work, or if they are in direct relation with the patient's occupation.	Never filled in.
I33	Where	Text	In case of a sudden event, the localisation of the patient at the time of the accident	Never filled in.
I34	Accident factor	Text	In case of a sudden event, the unexpected event or declination that provoked the accident.	Never filled in.
I35	Pulses (per sec)	n°	Vital sign	Never filled in.
I36	Breathing (per sec)	n°	Vital sign	Never filled in.
I37	Blood pressure	n°	Vital sign	Never filled in.
I38	Temperature, time	n°, time	Vital sign	Never filled in.
I39	Medical history (Anamnesis)	Text	The doctor's conclusions on the patient's previous diseases	Only information relevant with the present desease will be filled in.
I40	Symptoms	Text	The doctor's conclusions on the symptoms of the present medical trouble	As prescribed, plus vital signs etc., depending on the specific case
I41	Medical directions / medicaments	Text	Doctor's medical directions to the patient.	As prescribed, and always in the form of a list of directions
I42	Probable Diagnoses	Text	Doctor's probable diagnoses	Rarely filled in, only if the MD feels confident on his diagnoses.
I43	Decision type	Text	Treatment on board, need for monitoring and time of next medical contact, evacuation, etc.	As prescribed
I44	Evolution of patient's condition	Text	Progress, deterioration etc.	It is filled in at a later time; only when there is feed-back from the vessel.

A complete list of findings related to the real use of the MCF can be found in ANNEX II.

The most significant finding was that all fields concerning vital signs remained blank. Such information was considered important and indeed existed in many MCF records, but not in its prescribed location. It was rather incorporated as part of the symptoms story line in free text. For instance, instead of writing down patient's temperature in the predefined field, MDs incorporated such information on the symptoms field (Figure 9.5). Below there is a transcription of such a record.

> *"At 18:00 temp, O.K. In fact, it oscillates, the patient sweats and around midnight at 27+ h he felt pain at the fingertips..."*

It is important to note that during the interviews most MDs were arguing that the vital signs fields were important and that they were regularly used; however, the analysis of past records clearly contradicted their statements. In fact, noting such information in the free text "Symptoms" field had become a tacit convention among all MDs, but not explicit to their minds. In the interviews, MDs had a tendency to focus on the written result as a typical "Medical Record" and give little attention to its role during the teleconsultation process. This is probably because, the MDs were reflecting on their practice in procedural terms, whereas field observations led to the conclusion that the teleconsultation process was more driven by the specifics of the medical case at hand than by a predetermined medical procedure.


Example of the Medical call Form page2 (Translated in English. The numbers under parenthesis do not figure in the original MCF)

GENERAL MEDICAL INFORMATION AND VITAL SIGNS

WHEN(D/T):___/___(I30) HOW(S/P):________(I31)DURING WORK(Y/N):______(I32)

WHERE:__________ (I33) ACCIDENT FACTOR:____________ (I34)

PULSES:_____/SEC(I35) BREATHS:_____/SEC(I36) PRESSURE:____/____(I37)

TIME/TEMPERATURE: ___/___ ___/___ ___/___(I38)

PERSONALANAMNESIS:____________________(I39)

DESCRIPTION OF DECEASE / SYMPTOMS:__________

PROBABLE DIAGNOSES:________________(I42)

DECISION TYPE:________________(I43)

EVOLUTION OF PATIENTS CONDITION:__________(I44)

Project	Dissemination level	Page	Writer	Date	Vers	Document
ET-ASSIST / WP 02	P.P	93	[illegible]	28/04/2004	6a	D2.1 "Specifications of Remote [illegible] Aid Pilot"

Figure 9.5. Vital signs information incorporated inside the Symptoms free text field.

Another interesting finding was that the field “Probable Diagnoses” was also left blank. After asking MDs about this rather bizarre omission, their answer was that although the field was indeed important, in most cases they hesitated to fill it in because any stated “probable diagnoses” if errant, could potentially be counted as medical malpractice. What is even more interesting is that this information field had been renamed in the MCF three times in the past by the MDs themselves. Specifically, going through older versions of the MCF it was found that the field was initially named “diagnosis”, later revised to “probable diagnosis” and finally “Probable diagnoses”. In fact, these successive renaming indicated the MDs anxiety towards formally writing down in explicit terms a medical statement that could possibly invoke medical liability issues.

9.2.11 Evolutionary dynamics of the MMAC work practices

The above findings, among others, demonstrated that the actual practice of the Center was not in line with its officially declared procedures. Although this discrepancy between actual and formal is a usual and well-known phenomenon in large bureaucratic organizations, the puzzling thing in this particular field was that although the Centre was part of a large bureaucratic organization, and thus prone to such discrepancies, it nevertheless enjoyed a large independence in prescribing its functioning.

Indeed, the interviews with TOs and MDs revealed that MMAC personnel had almost total freedom to define and revise their operating procedures. For instance, it was found that the personnel had modified the MCFs format by their own initiative several times in the past (Nathanael and Marmaras, 2005). In 1987 when MMAC started its operation, it inherited along with the MDs, the MCF from the then National Medical Emergency call center. This initial form was tailored to the needs of the general public. Administrative information contained address, patient name and personal details, telephone number, etc. Medical information contained mostly vital parameters such as blood pressure, pulse rate, temperature, etc., along with a small descriptive field and a field for the decision taken (e.g., evacuate by ambulance).

As soon as in 1988, the members the center started to make modifications to this original form. Apart from obvious ones, such as the addition of fields like “vessel name” and “International Call Sign”, they also added a number of very specific fields such as “material transported”, “communication quality”, “pharmacy type”, “N° of Call”, etc. These adaptations did not happen at once. They emerged one by one, based on the collective accumulation of experience of MDs and TOs alike. For example, the field “material transported” was added after MDs realized that many poisoning cases in commercial vessels had to do with hazardous cargo. Initially, it started to appear on MDs side notes, and after a while it got formalized as a field, on the operator’s side of the MCF. In a similar manner, the field "previous call" was added after TOs realized that when searching in the archives for earlier MCFs of a particular case, they were never sure when to stop.

Most of the alterations concerned the TOs side of the form, whereas the MD side was kept almost identical to the original one. In fact, it was as if the administrative part of the MCF was evolving according to the needs and accumulated experience, whereas the medical part had stagnated, marking a clear discrepancy between formal information structure in the MCF and actual recording practice. What is more, this occurred with an astonishing constancy.

Nevertheless, this de-facto deviant use of the MCF had not been accepted by the MDs as such. The reasons behind this are not trivial. In fact, there had long existed quarrels among MDs concerning the quality of medical records, as some MDs tended to be less exhaustive in their records than others. Thus, at least in part, the stagnation of the medical side of the MCF might be attributed to the inability of the MDs as a team to come in terms on a harmonized recording practice.

Furthermore, in the interviews, MDs had a tendency to consider the MCF records as a typical "Patient Record" disengaged from the way it was produced and/or consulted in practice. Many were insisting on a highly structured form (e.g., with specified fields for every vital sign), whereas observations suggested that both the teleconsultation process and symptoms recording were more driven by the specifics of each case at hand than by a predetermined medical procedure. In short, MDs seemed to think of their everyday practice more from an academic standpoint than from a practitioner one, probably as a consequence of their formal medical training.

This contrast between what the MDs were claiming to be doing and what they were actually doing was evident in the specifications of a previous software application, commissioned by the Center in 1993. The study of the software's specifications documents showed that the MDs, in a marked contrast to their everyday practice at the time, had requested a digital form that contained more than 100 fields for medical information! The software was implemented and put to use, only to find out that it was almost impossible to work with. This was mainly due to MDs preference in using descriptive narratives and notes instead of predefined and structured information items, as this was more convenient and closer to traditional medical practice. During the interviews, the MDs noted that input of medical data on the computer was very laborious and in reality, it could only be done after the actual teleconsultation.

In fact, the system was conceived based on an idealized normative teleconsultation process, where all information required by the information system would be relevant and available at the moment needed. After a

6-month tryout, the software was put aside, and the Center resumed its operations with the old paper-based system, continuing small-scale modifications to the MCF format.

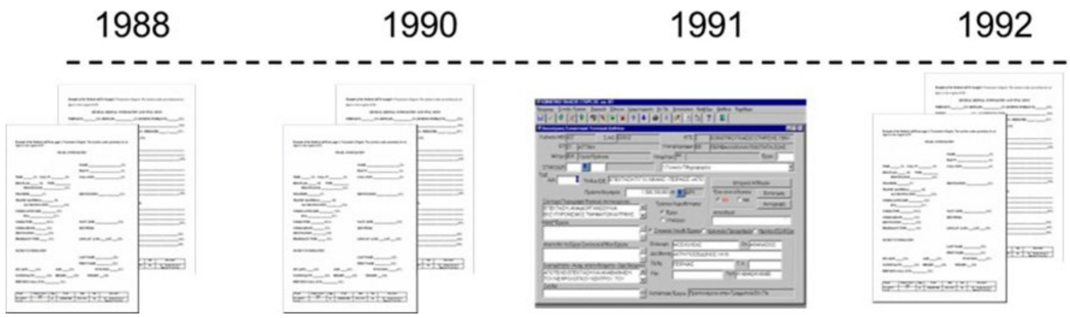


Figure 9.6. Historical progress and changes of the Medical Call Form.

A brief summary of the evolution of the MCF is the following (Figure 9.6):

- Between 1987 and 1992, progressive modifications to the MCF based on accumulated experience. Driven mostly by the pragmatics of everyday practice without a meta-reflection upon the whole process.
- In 1993 an effort to structure the MCF by means of information technology. From 23 administrative fields (at the time) and 15 medical ones, the new MCF contained 40 administrative and 100 medical fields. Sudden efforts to further codify information based on meta-reflection of MD's and driven by the perceived opportunities offered by information technology.
- In 1994, after a brief period of use, the new software, along with its codification, is abandoned. The new codification is judged too detailed for the task, and the software cumbersome and time consuming. The Center returns to the old paper-based MCF.
- Between 1994 and 1998, small scale modifications to the administrative part of the MCF to match practice needs (e.g., elimination of the "geographical coordinates" field, addition of a "shipping company telephone" field, etc.). No modifications to the medical part, albeit the mismatch of the MCF medical fields with the actual medical practice.

Actually, this evolution demonstrated a shift from a procedural-structured process to a more ad-hoc one. This shift had a positive effect on workload both in recording and in consulting MCFs and was more natural to the doctors' mental representations. Its main drawbacks were non-uniform records (varying among MDs), absence of definitive outcome or diagnosis of a case, unreliable search and difficulties in statistical analysis.

Regarding the TOs, the analysis of past records and prior versions of the MCF shed light to a number of recurring problems. One such problem was finding past MCFs for a specific case. TOs had tried two solutions in the past by adding specific fields in the MCFs, i.e., “N° of Call” in 1990 and “Previous Call Yes / N°” in 1997. None of them had worked in a satisfactory way; therefore, the TOs kept relying on the vessel’s interlocutor for such information. Obviously, this was an elementary problem to rectify through IT, but the identification of the above obsolete artefacts, gave an objective indication on the importance that should be granted to this task in the new tool.

The above work analysis combining field observations, past records, and ethnographic interviews permitted the designers to form an intimate understanding of the operational environment of MMAC. User needs were thus derived from an evidence-based analysis of the current work process and its historical evolution, and not through disengaged or typically wishful user requests. The above analysis permitted to identify both the main advantages and disadvantages of the current work process. Table 9.2 summarizes the identified positive and negative aspects of the current process.

Table 9.2. Positive and negative aspects of the current process identified during the user needs analysis.

Positive Aspects	Negative Aspects
- Flexible information recording through the use of paper & pencil - Reliable, robust process with no technological dependencies, posing minimum constraints to users. - MCF structure is easy to modify and can be modified by the users alone. - Information recording independent of telecommunication type with vessels.	- Actual practice does not necessarily follow the prescribed procedure. - Manual information recording does not facilitate statistical analysis. - Information on a case is fragmented (MCFs of a single case are stored separately) - Does not support automation possibilities (search for old MCFs is manual and time consuming) - Heavy dependence on user expertise -experience due to the informal nature of the process - Absence of standardization in the medical outcome / diagnosis

Following the User Centered Design philosophy, the central aim for the new system, was to preserve as much as possible the positive aspects of the current process while doing away as far as possible with the negative ones.

9.2.12 Comments of the need-finding process

Conducting such an analysis is not straightforward and, depending on the situation, can become tricky. Fiddling into past archives, unearthing discarded software documentation, consulting outdated procedures, etc., may seem strange to the people who pay you to help them design their future work. For most, it is not obvious why one should dig in the past in such detail to design a new tool. Besides, it is only natural for those who were responsible for prior ill-fated interventions, to hold a defensive stance towards their past decisions. Nevertheless, the analysis briefly presented above grounded on hard data, allowed the design team to identify a number of recurring problems and derive from them a set of invariants proper to the field of practice at hand.

For instance, the analysis made clear that the medical teleconsultation is a delicate cognitive activity that needs a support adapted to the accumulated work habits of MDs. However, any software design effort inherently stresses towards formalization, be it codification of information or proceduralization of the consultation process. The resistance of MDs in adopting all previous formalization efforts showed that the medical teleconsultation activity was clearly incompatible with such an approach. Thus, the new design needed to consider the MDs way of thinking and acting as invariant (i.e., a persistent constraint).

As a final note, many of the issues identified during the requirements analysis of MMAC, such as archiving MCFs by time of call and not by case, or the importance given by MDs to vital sign fields could be traced back to its establishment back in 1988. As MMAC was born from the remnants of a general public emergency call center, it inherited the practice of its predecessor. Thus, many of its operational specificities were, in fact, residues of the old center's practice (well adapted to the civilian public, but not to the maritime domain). Understanding this fact, as well as learning from the failures of all prior efforts to adapt the center's outdated practices to its new mission, allowed to form an understanding of its transformation dynamics.

Adopting such a transformation dynamics approach, tracing back the co-adaptation between MMAC and the various artefacts introduced to it in the past, allowed the design team to extend their horizon of observation, and thus, gain more insight on the center's possible attitude towards the current transformation effort.

9.3 User requirements specification

The user requirements were elaborated at a **second phase**, based on the findings of the user needs. As already discussed in Chapter 5 the transformation of user needs into user requirements results in a "wish-list" or "design brief" destined to drive main design choices. For the system in question a list of 18 application requirements statements and

12 interface requirements statements was elaborated. An indicative list of user requirements statements is given below.

9.3.1 System scope

The system shall support the following MMAC tasks: (i) real time recording of medical advice information to vessels, (ii) management of medical files (e.g., archiving, searching, and statistics), and (iii) provision of medical certificates.

9.3.2 System requirements (indicative)

1. The system shall be designed to assist the current operation practice without modifying the organizational structure and work processes.
2. The new process shall not increase the current mean telecommunication time with the vessels.
3. The information fields supported by the system shall be based on the ones actually used in the current Medical Call File (MCF).
4. The system shall support statistical analysis and presentation with minimum effort. The need for statistics shall not affect the doctor's way of recording medical information, nor the teleconsultation time.
5. In case of a system failure, the users shall be able to undertake their task manually.
6. The system shall be able to recognize an on-going case from a new one and present the existing relevant information on the case.
7. The system shall be able to recognize a vessel and/or a person that has received medical assistance by MMAC in the past, and present automatically all relevant information.
8. The system shall be able to support more than one patient in a single call.
9. The system shall recognize the MMAC users on every call (doctor and operator that handled the call).

9.3.3 Interface requirements (indicative)

1. The interface shall support the operational sequence of a teleconsultation with minimum screen changes.
2. The interface shall be designed in a way that learning time would

not exceed two working days for users, not familiarized with computers.

3. Administrative data entry shall require the least typing possible from the part of the TO.
4. Medical data entry shall pose minimum constraints to MDs, as this type of activity greatly influences their heavy parallel medical (cognitive) task. Due to low familiarization of the MDs with the keyboard a graphics tablet was mentioned as possible solution. The foreseen solution shall minimize the risk of system rejection by the MDs.
5. During the tele-consultation period, the TO shall be able to view all previous administrative entries and change most of them.
6. During the tele-consultation period, the MD shall be able to view all previous medical entries on the case without changing screens.
7. During the tele-consultation period, the MD shall be able to view a part of administrative information without changing screens.
8. Statistical data entry of a case shall require no typing and need minimum time and effort to be completed by a trained MD.

9.3.4 Rationale for the main design choices

Considering (i) MMAC management's explicit demand for a software tool, and also (ii) the prior effort to develop such a tool, back in 1993, as well as (iii) the MDs insistence on keeping such fields in the MCF (without actually using them), dictated that there was indeed a need to formalize parts of the process. If it could not be the teleconsultation activity, then it could be a summary of its results. The above led the design team to propose a decomposition of medical information input in two distinct phases. One almost totally free of formalization, destined to support the teleconsultation sessions in real time (possibly by keeping MDs handwritten notes), and a second one, destined to summarize the results of all sessions in a codified way at the closure of each medical case. The idea of decomposing medical information in "active – unstructured" and "summarizing – structured", came directly from the identified contradiction between the MDs need for freedom during teleconsultation, and their need for structured summary at the end of each medical case.

However, even a codification of the medical summary presented challenges. As noted above, the evidence assembled during the analysis of past MCF records and prior versions of the MCF, suggested that MDs were reluctant in writing-down strong diagnostic statements. This was evidenced by the "probable diagnoses" field being left blank, but also from the fact that its label had been modified three times in subsequent versions of the MCF, towards more supple wordings (from “Diagnosis” in 1988 to “Probable Diagnosis” in 1994 and to “Probable Diagnoses” in 1997).

These findings led the design team towards proposing a flexible solution for the codification of medical problems, based on a simplified scalable version of ICD 10 (International Classification of Diseases – Revision 10) to be implemented through a three-step hierarchical selection menu. This enabled the MDs to be as specific as they felt they could be, each time, thus alleviating medical liability risks.

9.4 Functional system design

The **third phase** in the design process was the functional system design. This included the definition of data, database structure and system workflows for the two workstations (TO and MD). System functionality was designed under three principles:

- coherency
- flexibility
- visibility

Furthermore, processes were designed to take into account the current work organization and the unofficial work practices. Care was specifically taken so that the new processes reflect the users’ representation of the work as much as possible. The main functions supported are summarized below:

- processing of calls,
- access authorization,
- archiving – searching – maintenance,
- case closing,
- issuing of certificate,
- statistics generation / presentation,
- system maintenance.

Meetings of the design team with user representatives were conducted to discuss and conclude on dataset definition and structure. The effort was put on attenuating as far as possible the natural user tendency for over-specifying information. To this end empirical evidence from past medical records played a pivotal role as a basis for the new dataset and structure. Finally, from the 30 information fields contained on the TOs side of paper MCFs, 23 were retained in the digital version, and from the 15 information fields on the MDs side of the paper form, only three (3) fields were retained for supporting the medical teleconsultation and another three (3) multiple selection fields for formal documentation of the case outcome.

Data structure was also defined in collaboration with user representatives. At a first phase, a number of ambiguities had to be resolved at the conceptual level, since MDs had diverging views on the definition of various entities (i.e., some MDs related a case with only one patient whereas others maintained that a case could entail more than one individual). For this purpose, a user-friendly representation of the conceptual data structure was produced as a means for discussion and facilitation of consensus among users (Figure 9.7).

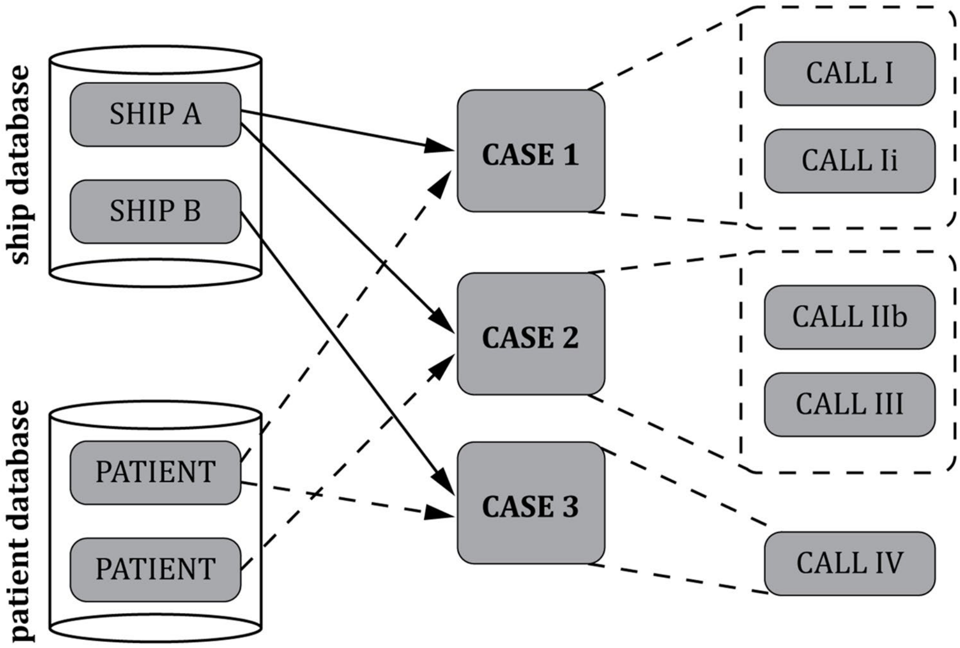


Figure 9.7. User friendly representation of data structure for discussion with users and arriving to a consensus among them.

The detailed database structure was, then, elaborated by the software development team, based on the rudimentary conceptual data structure agreed and validated by uses (Figure 9.8). Involvement of users in this process was pivotal for both technical reasons, i.e., (i) to enable a coherent mental model of the data structure among users, and (ii) to render this mental model compatible with the technical data model, as well as for motivational reasons, i.e. (iii) to enhance the sense of participation / ownership and therefore acceptance of the new system from the initial design phases of the project.

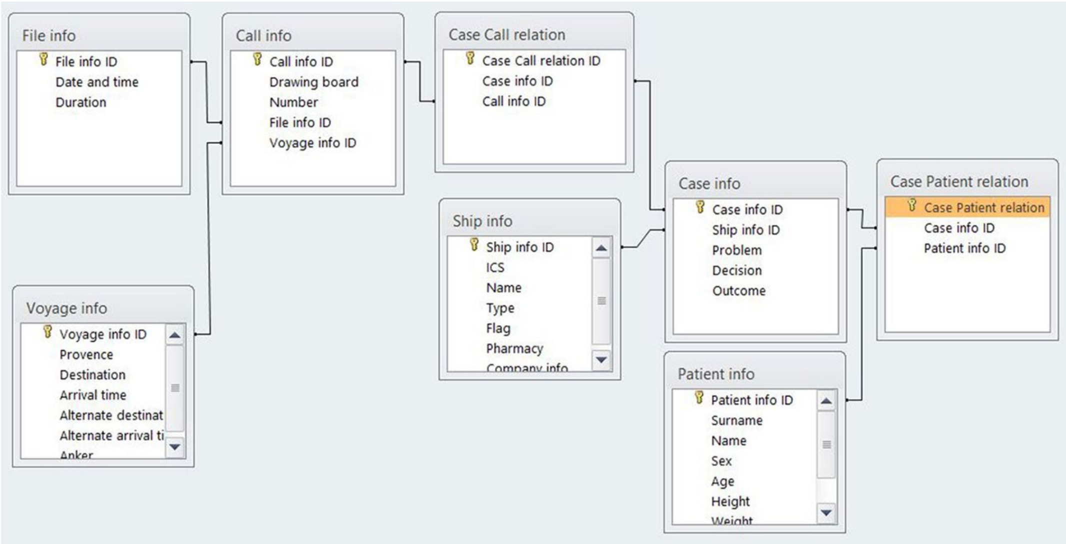


Figure 9.8. Overview of implemented Data Base structure.

Teleconsultation workflow was specified first by means of a flowchart (Figure 9.9), so as to enable user's understanding and acceptance of the proposed task allocation among TOs, MDs and system. Care was taken at this stage in order to allow dynamic adaptation of the workflow depending on the circumstances in each call (e.g., passing a call directly to MD before filling-in administrative details in cases of extreme emergency). All other workflows were specified and discussed with users in a similar fashion and, after agreement, were forwarded to the development team.

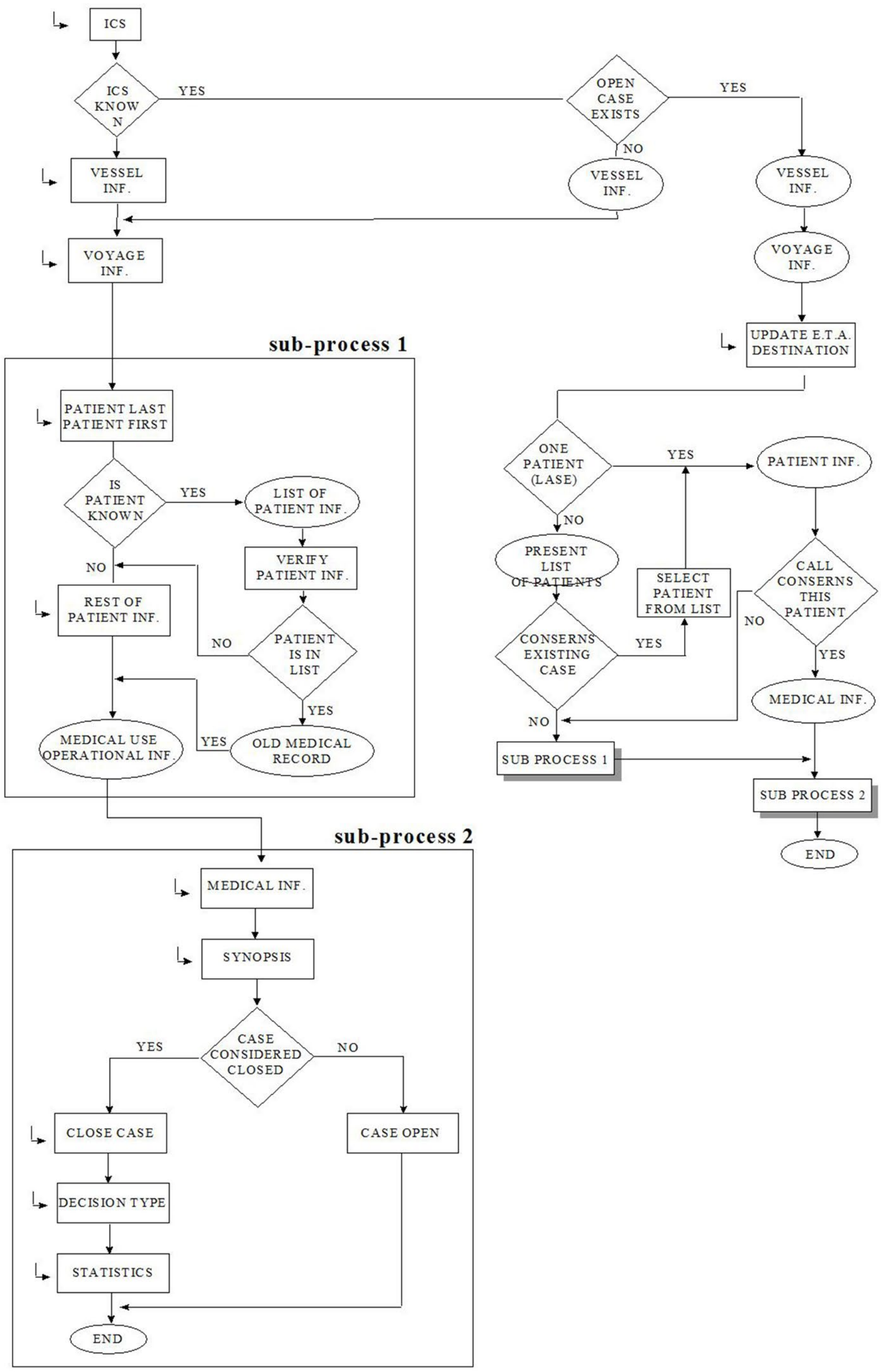


Figure 9.9. Flowchart of the Teleconsultation workflow: these diagrams proved useful for discussing design alternatives with the prospective users.

9.4.1 User interface Design

The **fourth phase**, i.e., the design of the user interface, followed an iterative approach starting with paper-based sketches and followed by various versions of digital mock-ups. User representatives (one TO and one MD) were directly involved by participating in two half day ideation workshops.

The design of TO workspace was rather straightforward. Good semantic clustering of information (e.g., List of Cases, Vessel, Voyage, Communication and Patient inf.) plus sequential dependencies on filling-in information drove the design choices. Input means were keyboard and mouse. One of the early digital mock-ups of the TO graphical workspace is shown in Figure 9.10.

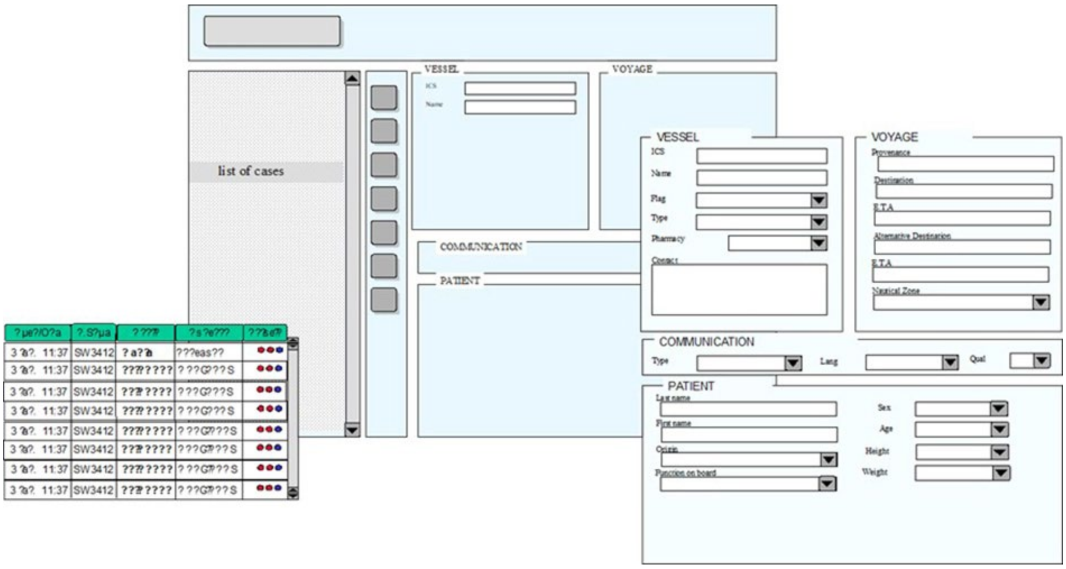


Figure 9.10. An early digital mock-up of the TO graphical workspace showing details of active cases list (below left), general screen layout (center) and information field segmentation (below right).

For the design of the MDs workspace there were several challenges. A main issue was the input means. MDs were very reluctant to change their pen and pencil way of noting down information while in a teleconsultation session. In fact, most MDs at the time in MMAC had very limited computer experience, and it was clear that forcing them to use a keyboard for medical notetaking would almost certainly result in double record keeping (i.e., one handwritten on paper in real-time and a one typed into the system a-posteriori). Note that at the time OCR algorithms for handwriting were unreliable. Therefore, it was decided to experiment with a digital handwriting solution with the use of a graphics tablet / digital pen.

Another issue concerned screen real-estate. While on teleconsultation mode, MDs not only had to write down notes but also to be able to consult (i) administrative information entered by the TO, and (ii) prior medical information entered by another MD in prior call(s) of the same case. Based on detailed observation data and MD opinions, a hybrid solution was chosen with administrative information always visible on the MD workspace and earlier calls easily accessed through tabs.

To answer to the above challenges, the final digital mock-up adopted a novel shared workspace concept. Specifically, a single visual workspace was provided to both TO and MD with dedicated intervention zones for each one (Figure 9.11). The left side of the workspace was reserved for administrative information, where TOs and MDs had full rights, while the right side was reserved to medical notes where TOs could, also, observe but not intervene. In this way, both team members remained engaged in the teleconsultation process, fostering mutual visibility and flexibility in information input and viewing. The same basic workspace was retained for all other system states in the mock-ups (e.g., list of cases screen, medical statistics screen), in an effort to retain simplicity in navigation and transparency at all times.

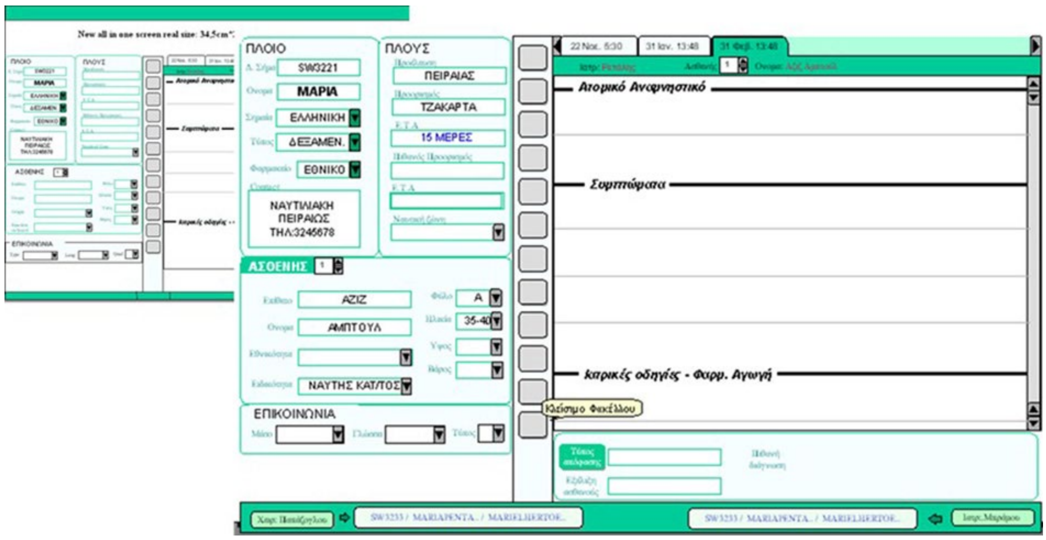


Figure 9.11. The final digital mock-up of the teleconsultation workspace. The left side is dedicated to administrative information while the right one to medical notes. Tabs on the medical side signify note of subsequent calls of the same medial case. The middle column incorporated the main menu buttons so as to be easily accessible by both the TOs mouse and MDs digital pen.

9.4.2 Menu button selection

Users were, also, involved in the selection of most appropriate icons for main menu buttons. Three alternative icons were proposed for each action button, and all users were called upon to choose the icon that they felt better conveyed the intended action button meaning. The form used for icon selection, as well as the three alternative icons for the new medical call button, are presented in Figure 9.12. An icon for an action button was retained if chosen by 60% of users or more. Otherwise, new alternative icons were proposed until user choices converged.

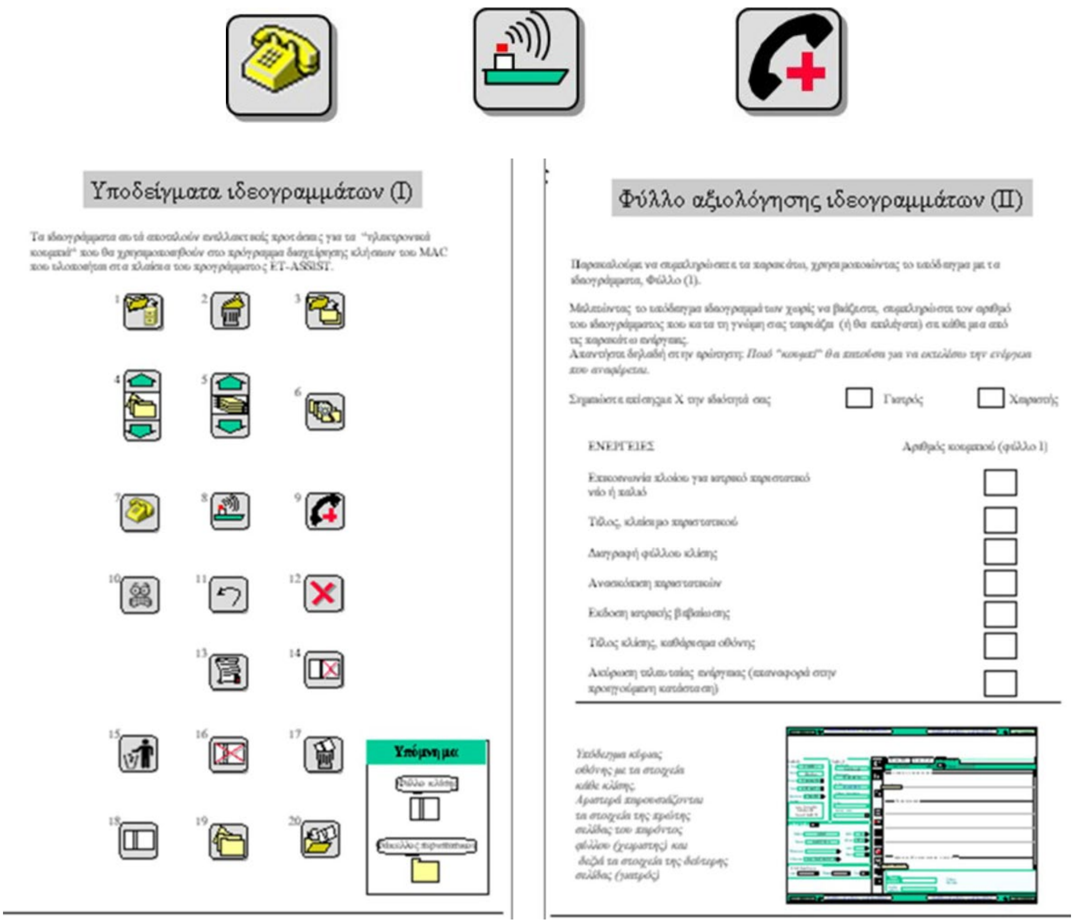


Figure 9.12. Alternative icons for the new medical call (top); the form used for the icon selection poll (bottom). Both in Greek language

9.4.3 Navigational structure

At a next stage the navigational structure among system states was specified through a State Transition Diagram (Figure 9.13).

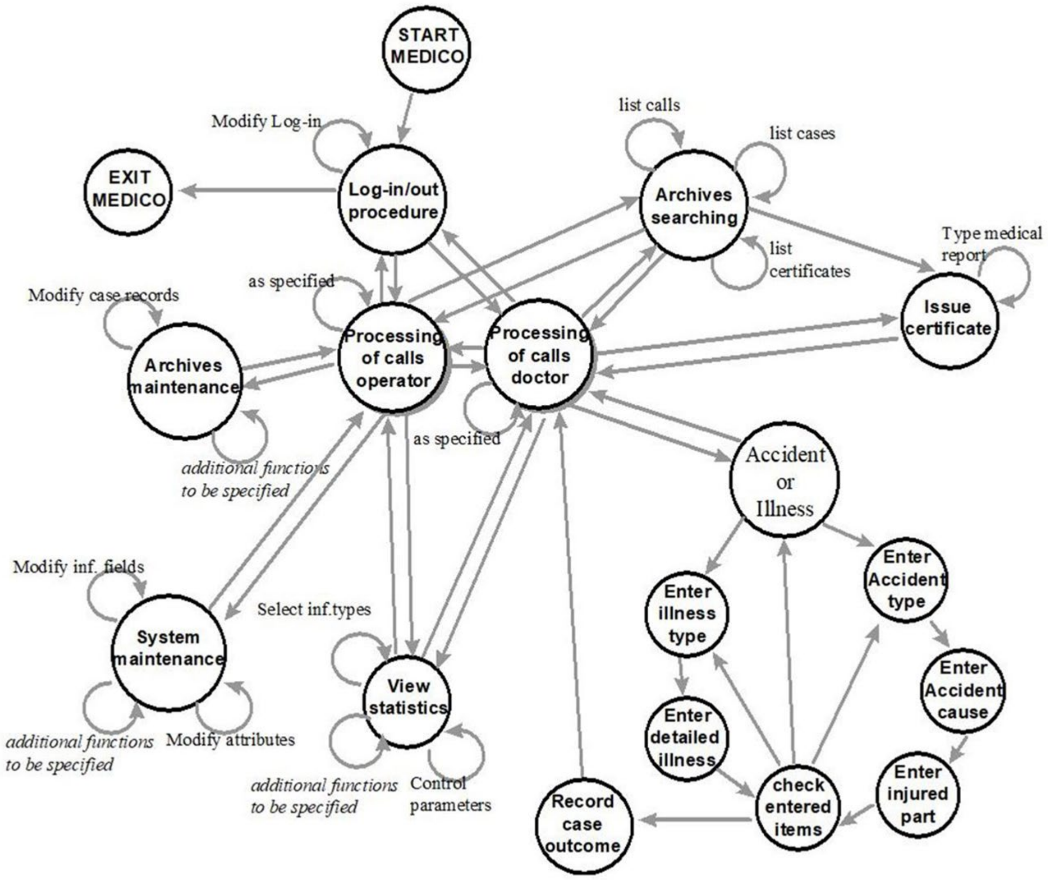


Figure 9.13. Representation of possible system transitions between states. Each state (derived from the processes specified above) is represented by a cycle and signifies a different screen (or screen class). Each transition is represented by an arrow and signifies actions resulting in screen (or state) changes.

The final mock-up and state transition diagram were reviewed by a representative number of users (MDs and TOs) to conclude on various details. The resulting design was used as part of the final system specifications.

9.5 Prototype development

At the **fifth phase** of the User-Centered Design process, a prototype of the system was developed. The development was done in Visual C++ to allow for maximum freedom in the graphical interface design. The first functional prototype was named MEDICO, after the distress signal “pan-pan medico” used for medical incidents in voice-procedure radio communications onboard ships.

MEDICO had the following features:

Two workstations with a single virtual workspace environment for both MDs and TOs tasks (Figure 9.14); the left part was reserved for administrative data while the right part was reserved for medical data. Thus, both users worked on the same virtual workspace from different physical workstations.

Different types of input means adapted to each task. The TO used a keyboard and mouse. The MD used a graphics tablet with a digital pen; medical input was done by handwriting or by simple selection. In fact, the medical part of the screen may be described as a hyper-notebook, having the familiar pen-on-paper feel, while offering extended features such as effortless erasing, copy and paste, search between pages, data entry by selection, etc. In case a MD wanted to intervene to the administrative data, the pen could be used for this purpose through a virtual pop-up keyboard, by means of double-tapping on alphanumeric fields.

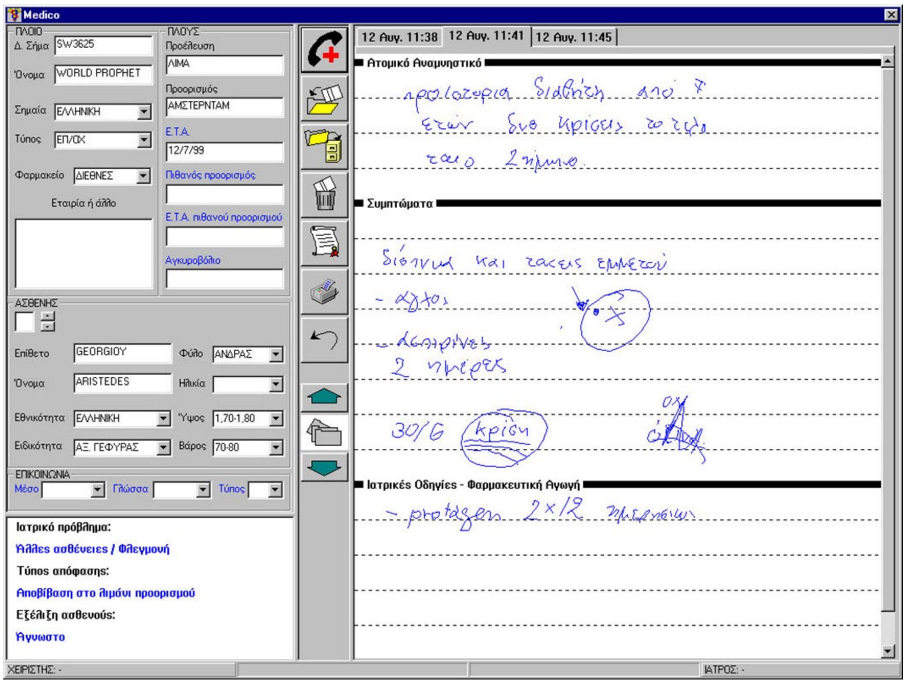


Figure 9.14. MEDICO's virtual workspace during tele-consultation

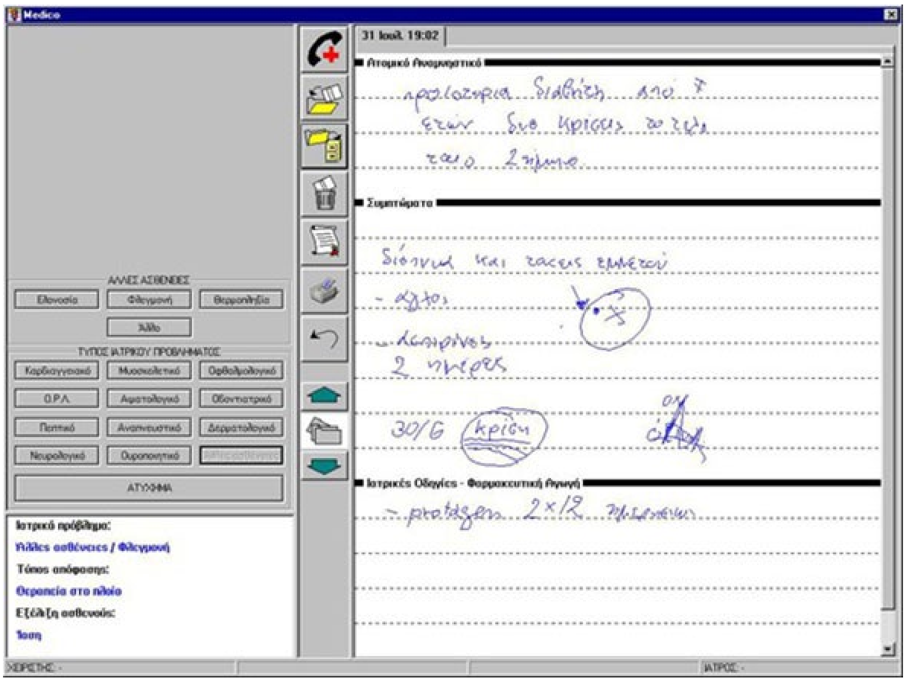


Figure 9.15. MEDICO's workspace during input of medical data for case outcome. The left side of administrative data (of no use in this task) is replaced by a multiple selection menu (derived from ICD-10), while the right side with medical notes on tabs remains visible.

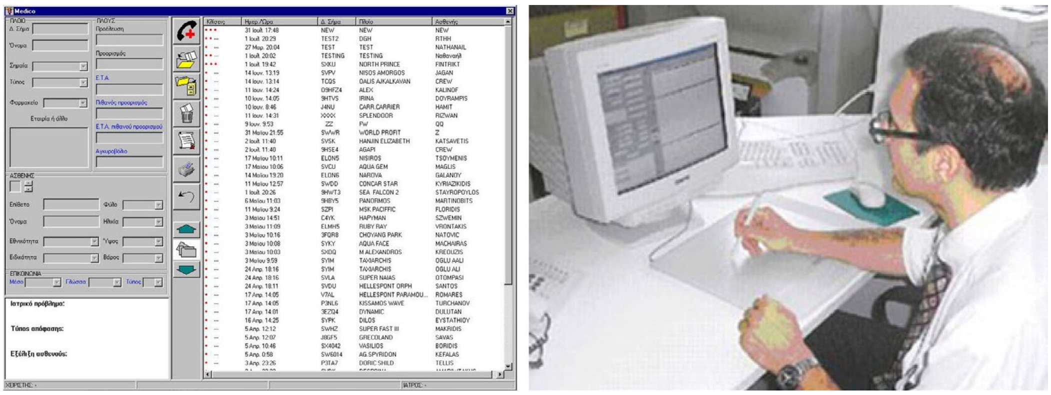


Figure 9.16. MEDICO's workspace in idle mode presenting the list of active cases (left); an MD of the center testing the handwriting feature of the prototype (right)

Division of medical information input in two distinct stages; (i) operational medical information, which is used as MD's memory support during the tele-consultation (Figure 9.14) and (ii) medical information which is used for case outcome documentation and statistics (Figure 9.15). The first is entered by handwriting and saved as descriptive medical records (notes). This way of data input is perfectly adapted to MDs operational practice, permitting simultaneous notes-taking while speaking on the telephone with ships. The second is entered by multiple selections through predefined menus, achieving the required standardization for statistical analysis. This standardized information is entered after a particular case is over, since it is only then that it becomes definite.

Grouping all calls of a medical case and displaying them on the same screen. Administrative data for all calls of the same case remain stable, while medical data of different calls are displayed by changing the tabs of the hyper-notebook.

Centralized control of system's functions from the same screen. When the system does not display a particular case, the hyper-notebook (left part of the screen) becomes a list of medical cases with sorting facility (Figure 9.16, left). When a doctor decides to close a case, the administrative information (right part of the screen) displays the medical statistics menu. In this way, the system achieves optimum use of the display space adapting itself according to the particular task.

Well-structured authentication and read – write rights. For instance, medical entries can only be modified by their author, while administrative ones can be modified by any user. Moreover, the system provides a memory aid by reminding to particular MDs to close cases that have been inactive for more than 15 days.

9.6 Evaluation of the MEDICO prototype

The MEDICO prototype was evaluated at the **sixth phase** of the User-Centered Design process through user group meetings, training sessions, direct observation of use. After a number of design improvement iterations, the system was introduced to the operations center. A problem log was kept for the first month of operation for validation purposes.

9.6.1 User group meetings

User group meetings were performed throughout the pilot phase to gather individual views of the users. Additional one-to-one meeting with all users in the form of training sessions proved to be very effective in evaluating the time and effort required to learn the system. This ranged from one hour for TOs to one workday for MDs (having no prior computer literacy). Reluctance to use a computer was the main problem with a number of MDs. This problem was well identified in the user requirements phase, and thus, the handwriting solution that was chosen of for MDs through the use a graphics tablet instead of typing effectively overcame the "computer fear syndrome".

9.6.2 Direct observation techniques

Direct observations of system use were performed for three work shifts (a total of 24 hours), both with past and real-time cases. Results showed that there were significant improvements in retrieving and communicating medical case data, without negatively affecting the teleconsultation process (a summary of quantifiable operational efficiency impact of the system is presented in Table 9.3).

Table 9.3. Summary of overall operational efficiency impact

Measure	**Impact**
MCF retrieval time	from 30 sec (mean time) to 2 sec
Archiving time	from 30 sec (mean time) to 2 sec
Administrative inf. Input	no significant change
Medical inf. Input	no significant change
Administrative inf. Completeness	from 30% (mean) to 70%
Medical outcome inf. Completeness	from 40% to 90%

However, observations also revealed that MDs flexibility in scheduling their work was narrowed (e.g., taking a medical call on their mobile phone in cases of temporary absence). Since the user requirements stated that the system should have the least possible organizational impact on the center's activities, alterations on the system were made to deal with this problem.

9.6.3 Problem log analysis

A problem log was kept by the system operators throughout the validation period to record any remaining technical or operational malfunctions with the use of the system. The great majority of problems recorded were technical in nature (because of no sufficient time for full debugging). Apart from the above, the system fully supported the operational needs of the users after one month and 100 cases, which were managed through the system. Overall, the evaluation process concluded that the new information management system (MEDICO) fully supported the operational needs of MMAC and was welcome by the users.

9.7 Discussion

The basic philosophy adopted while designing the system was to preserve the positive aspects of the earlier work practice, while doing away with the negative ones. Special effort was put to preserve as much as possible the flexible nature of everyday practice and the actual organizational structure.

A direct consequence of this philosophy was the use of handwriting for the recording of medical information. Furthermore, complying with the previous practice, both medical and administrative tasks are carried out through a single screen environment. Furthermore, the system did not feature unnecessary functionality. Information technology solutions were driven exclusively by the user requirements, resulting to a system's image that is simple, practical and compatible with the users' representation of their tasks.

The system designed was efficient, effective and easy to learn. Moreover, the above main features of the system had a positive effect on its acceptance by the users. Indeed, the migration period lasted only one week with minimal training.

Some lessons learned from this project are the following. Typically, the main effort in IT systems development is put in supporting as many working scenarios as possible, often to the detriment of system's usability. Users, as a rule, press for more functions than they actually need. Last but not least, developers, in order to add value to their work, tend to implement all the functions that are easily supported by the platform they use. In the present project, these two apparently conflicting requirements (support as many working scenarios as possible versus system's simplicity and usability) were met by insisting on the results of the thorough work analysis of the earlier work situation. This permitted the formalization of explicit and unambiguous user requirements, which have led to precise and elegant design solutions ensuring high usability.

The clear decision to preserve the nature of the task and to minimize the system's effects to operational processes and the organizational structure, allowed for a seamless implementation. Concluding, it is important to note that the overall project management was carried out by ergonomists. This ensured an authentic user-centered design and prevented the system from slipping towards technology-driven "cumbersome" solutions during the development period.

As a concluding remark, a user-centered approach in the design of interactive systems puts the future users and context of use at the center of the design process. The aim is to achieve excellent user and task support within a given context, rather than expecting users to alter behaviors to accommodate the system. In User Centered Design (UCD), a deep understanding of the work domain requirements and constraints, as well as the way users cope with them, come to the fore of the design process before technological choices are made. A central imperative of UCD process is, therefore, the direct involvement of future users in all design phases.

References

Nathanael, D., & Marmaras, N. (2000). User-Centered Design in Practice: A Medical Tele-Consultation Management Application. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44(2), 346–349*. https://doi.org/10.1177/154193120004400214

Nathanael, D., & Marmaras, N. (2005, September). Historical analysis as a means to uncover the dynamics of evolving practices. *Proceedings of the 2005 annual conference on European association of cognitive ergonomics* (pp. 65-70).

ANNEX I

Extract from the observation notes at the need-finding phase of MEDICO.

A typical medical case usually involves more than one calls to the Centre. Once a ship calls for medical advice, the call is picked up by the Operator. The standard process is the following.

The operator asks the calling ship if their call concerns a new medical case or an ongoing one. If the call concerns a new medical case, he asks for a number of administrative information on the Ship the Voyage and the Patient. If the call concerns an ongoing case the operator updates only the voyage information.	*The administrative inf. in the Medical Call Form (MCF) has been categorized in groups that reflect real world entities as the ship, the patient, the voyage and the call. This grouping adds structure to the operator's representation of the situation helping him ask and the caller answer the questions posed by focusing every time on human explicit entities.* *The questions of the operator are constrained by the fields in the MCF. The operator never asks for inf. that does not appear in the MCF (side notes are non-existent in the MCFs reviewed). On the opposite side only few fields are mandatory in the sense that they appear in all MCFs reviewed. (Formally all fields are mandatory)*
For a new medical case the operator records the information progressively in the MCF. He may be selective on the information asked according to the particular situation.	*The order of recording is partially structured by the configuration of the MCF. The MD is in most cases on the vicinity trying to get inf. On the call. MDs have been observed to overlook the information recorded on the MCF*
For an ongoing case, the operator after updating the voyage inf. asks the caller for the date and time of the last call and subsequently searches in the Case files for the MCFs of the previous calls on the case.	*The MCFs are stored in a folder in strict chronological order. The operator in order to facilitate his search uses the caller's memory of the date of last call as a hint. Absence of this information greatly increases search time.*

The Operator passes the call to the MD, passing over the MCF as well. He may verbally brief the MD any significant details on the call.	*The MCF is a double sided A4 form. One page for administrative inf. and the other for Medical. MDs have been frequently observed turning to the front page of the form to consult administrative inf.*
In most cases the MD speaks with a ship officer and not the patient himself	*The ship officer plays a mediating role partially rationalizing the demand. This filtering has both positive and negative effects on the inf. passed. (this filtering effect is explicit to MDs)*
If it is an ongoing case the MD quickly scans the previous MCFs, while listening to the caller.	*When there is more than one MCF the MD is trying to define their chronological order. The operator may prepare the MCFs in chronological order, thus adding structure to the MDs information gathering task.*
MDs use a complex approach in order to gain understanding of the medical problem listening to the complaint first and progressively interrupting to redirect the conversation to meaningful to them inf. The MDs hold scattered, unstructured notes that tend to occupy the two-dimensional space of the free text field in a fussy way rather than follow a typical linear writing style.	*The communication process is constrained by the medium (voice quality – doublex or non doublex channel)* *The conversation process is not affected by the MCF structure as in the case of the operator. The MCF is used as a memo pad for significant details that arise through the conversation. (even when dedicated inf. fields exist, e.g "temp." or "pulse", this information on symptoms is not recorded in the dedicated space in the MCF but in the general "notes" space reflecting the course of the consultation).*
Once the MD has formed an "adequate understanding" of the medical problem he proceeds by giving a series of "medical directions" to the caller. These directions are written down in a numbered order by the MD in real time in a dedicated section of the MCF called "Medical directions".	*"Medical directions" tend to be explicitly recorded and numbered in contrast to other medical inf. e.g. symptoms. (there are MCFs, where only the "medical directions" part was completed by the MD). Possible reasons are "what can harm is a wrong act upon the patient not an omission" or "ensure that if another MD is involved in the case later on, his directions (i.e., what is observable by the caller) are compatible with the previous ones.*

At the end of a call the MD may advise the caller to re-contact the Center at either after a specified period or upon a specific condition.	*The Center has no official means for contacting ships. So, a new medical consultation can only be triggered by the ship. The MD constrains the conditions for this triggering and records them as the final medical direction. The responsibility to re-contact is thus fully carried by the ship. MDs rarely reflect on the evolution of a case and then only if they have to consult an expert (i.e., continue their previous medical act). A notifying system located in the Center would not provide any benefit under the current telecom constraints.*
After the communication with the ship has ended the MD may record some form of hypothesis on a diagnosis or list the most pertinent symptoms of the patient.	*This inf. is written in the lower part of the "notes space". The dedicated field called "diagnosis" is never filled in.*

Additional notes

The perception of the medical activity by the MDs

The MDs are very sensitive in the terminology they use to describe their activity. They employ the terms “advice” instead of “consultation”, “adequate understanding” instead of “diagnosis” and “medical Directions” instead of “treatment”. This is most probably due to the feeling of uncertainty that they have concerning their medical activity. It may be partly accounted to a collective standing over liability issues and partly to an explicit recognition of the special character of their medical activity. It is interesting to note that the official medical certificates provided by the center (if requested) rarely include a diagnosis, they usually provide only a description of the circumstances, the symptoms and medical directions given.

The issuing of a medical certificate involves the production of an official document (a document that will circulate outside of the work system and may by itself influence the decisions of other organizations). Thus, if a certificate is requested, MDs may spend a considerable amount of time to formulate a coherent description (a post rationalization) of the situation and at the same time, paying attention not to over specify a situation which will always remain with some ambiguity.

Official work schedules & real practice

According to the formal procedure, there should be one operator and one MD at the center on a 24-hour basis. Shifts change every 8 hours i.e., at 06 am at 14 pm and at 22 pm. In every day practice a variability is introduced by two types of events.

An employee being late for work, resulting in his colleague of the previous shift staying in position waiting to be dismissed, and an employee leaving early from work or leaving before his colleague of the next shift has arrived.

The first type of event does not pose a problem in the service provided; (the rotary shift system ensures that employees that tend to be late will quickly find themselves in the uneasy position of waiting to be changed by an unhappy colleague, so equilibrium is achieved).

The second type of event is only observed in MDs and does have a negative influence on the service provided. For various reasons (most probably related to the power structure of the organization), some MDs will not always be physically present in the center and will attend their shift by roaming the calls to a cell phone. (By not being physically present, an MD cannot effectively collaborate with the operator and more importantly, they cannot consult previous MCFs on a case). Interviews with the management of the center revealed that this is an aging problem that they would like to tackle but have been unsuccessful up to now.

ANNEX II

A complete list of findings related to the real use of the MCF.

Information gathered by HRC MAC during the processing of a medical case included 44 information fields of which 29 were administrative and logistic information (thus, gathered by the operator) and 15 fields were medical information (gathered by the doctor). The following table presents a complete list of the information fields incorporated in the MCF accompanied by the results of the actual use of these information fields after detailed analysis of 500 past MCF records.

	Information field	Data type	Description	Actual use
Vessel information				
I1	MCF number	Month/Day/n°	This is the identification code of an MCF; It consists of the Month, Day and sequence n° on that date. (it is independent of the case)	As prescribed. (Case identifier is date of last contact plus vessel name)
I2	Vessel's name	Character set (vessels name)	The vessels registered name in Greek or Latin depending on the name's origin	As prescribed. (Spelling of vessels name may differ between different MCFs)
I3	Time of call	Hour/min	Athens local time at the moment of call reception	As prescribed
I4	Number of call	n°	This number is case dependent. It represents the order of calls in a case (1st, 2nd, 3rd call)	As prescribed. The operator has to find the previous MCF to validate the n°.
I5	International Call sign	Character set (code containing letters and numbers)	This is the identification code of the vessel and is unique as opposed to vessel name. The first digits give an indication of the registered flag.	As prescribed
I6	Registered flag	Character set (Country name)	Registered flag is the country where the vessel is matriculated	As prescribed

	Information field	Data type	Description	Actual use
I7	Weather	Text	Current weather conditions in the vessel's sailing waters	Never filled in
I8	Vessels type	Text	Description of the vessels type (e.g., bulk carrier, tanker, fishing boat etc.)	As prescribed
I9	Transported material	Text	The material transported, for commercial vessels only.	Almost never filled in. If judged relevant with the desease it will be recorded in the symptoms field
I10	Communication language	Character set (Language)	The language in which the Tele-consultation takes place.	As prescribed
I11	Communication quality	n° (1-5)	A subjective estimation of the operator on the quality of the voice link (takes values of 1 'poor' to 5 'excellent')	It is filled in, only in the case of a mediocre radio connection
I12	Communication means	Character set	The connection type. Can be INMARSAT, Athens - Hellas Radio, Telex or Fax	As prescribed
I13	Pharmacy type	Character set	Pharmacy type is the official pharmacy category. In practice it can be either International 'more complete' or national 'less complete'	As prescribed, although the pharmacy category gives little information on the actual pharmaceuticals on board
I14	Provenance	Character set (Country or port name)	Port or country from which the vessel started the present voyage.	As prescribed. The port is usually indicated when the voyage is inside Greece.

	Information field	Data type	Description	Actual use
I15	Destination	Character set (Country or port name)	Port or Country of destination of the vessel's present voyage.	As prescribed. The port is usually indicated when the voyage is inside Greece
I16	Alternative Destination	Character set (Country or port name)	Possible stop-over port in case there is a need for a patient evacuation.	It is filled in only in the case the doctor recommends a derouting for evacuation
I17	Estimated Time of Arrival	Month/Day/ Hour	The scheduled date/ time of arrival of the vessel to its destination. It is calculated daily by the vessel's navigation officer.	In practice what is recorded is the estimated remaining duration of the voyage
I18	Nautical Zone	Character set	The standard nautical code for the sea region the vessel is currently sailing in.	Never filled in.
I19	Distance from destination	n° (N.miles)	The present distance of the vessel from the destination port. It is calculated daily by the vessel's navigation officer.	Never filled in.
I20	Vessel position LOG/LAT.	n° (two four-digit numbers)	The geographical co-ordinates of the vessel's current position.	Almost never filled in. It may be useful only by rescue authorities in case of an emergency evacuation.
I21	Previous call	Y/N	Previous call designates whether there has been a previous call for the same case. This field is closely related to the field 'Number of calls.	In practice it is used to distinguish 2 medical cases in the same vessel

	Information field	Data type	Description	Actual use
Patients information				
I22	Last name	Character set	The patient's last name in Greek or Latin characters depending on the name's origin.	As prescribed. Spelling may differ between different MCFs
I23	First name	Character set	The patient's first name in Greek or Latin characters depending on the name origin.	As prescribed. Spelling may differ between different MCFs
I24	Sex	M/F	The patient's sex	As prescribed
I25	Age	n°	The patient's age	As prescribed (sometimes only a rough estimation)
I26	Nationality	Character set (Country name)	The patients nationality	As prescribed, although what is seeked is the patients race.
I27	Function	Text (patients occupation)	the patient's formal status on board the vessel. (e.g., Captain, bridge or engine officer, seaman, cook, passenger, etc.)	As prescribed
I28	Height	n°	The estimated patient's height	As prescribed
I29	Weight	n°	The estimated patient's weight.	As prescribed
General medical information and vital signs				
I30	When troubles started	Text	The temporal span of the medical troubles (the patient's estimation)	Never filled in. This type of information is routinely recorded in the symptoms field.
I31	How troubles started	Text (suddenly, progressively, after a... etc.)	The patient's perception on the aggravation of the medical troubles	Never filled in. This type of information is routinely recorded in the symptoms field.

	Information field	Data type	Description	Actual use
I32	During work	Y/N	If the troubles started during work, or if they are in direct relation with the patient's occupation.	Never filled in. If judged relevant with the desease it will be recorded in the symptoms field
I33	Where	Text	In case of a sudden event, the localisation of the patient at the time of the accident	Never filled in. If judged relevant with the desease it will be recorded in the symptoms field
I34	Accident factor	Text	In case of a sudden event, the unexpected event that provoked the accident.	Never filled in. If judged relevant with the desease it will be recorded in the symptoms field
I35	Pulses (per sec)	n°	Vital sign	Never filled in. If judged relevant with the desease it will be recorded in the symptoms field
I36	Breathing (per sec)	n°	Vital sign	Never filled in. If judged relevant with the desease it will be recorded in the symptoms field
I37	Blood pressure	n°	Vital sign	Never filled in. if judged relevant with the desease it will be recorded in the symptoms field
I38	Temperature, time	n°, text	Vital sign, plus the time the temperature was taken.	Never filled in. if judged relevant with the desease it will be recorded in the symptoms field
I39	Medical history (Anamnesis)	Text	The doctor's conclusions on the patient's previous diseases	Only information relevant with the present desease will be filled in.

	Information field	Data type	Description	Actual use
Description of the desease				
I40	Symptoms	Text	The doctor's conclusions on the symptoms of the present medical trouble	As prescribed, plus additional relevant information as vital signs etc., depending on the specific case
I41	Medical directions / treatment / medicaments	Text	Doctor's medical directions to the patient.	As prescribed, and always in the form of a list
I42	Probable Diagnosis	Text	Doctor's probable diagnosis	It will be filled in only when the doctor feels confident on his diagnosis.
I43	Decision type	Text	Decision type such as: treatment on board, need for monitoring and time of next medical contact, evacuation, etc.	As prescribed
I44	Evolution of patients condition	Text	Progress, deterioration etc.	It is filled in at a later time, and only when there is feedback from the vessel.

About the authors

Dimitris Nathanael is Associate Professor of Ergonomics and Cognitive Systems Design and head of the Ergonomics lab at the National Technical University of Athens, Dept. of Mechanical Engineering where he teaches: Ergonomics / Human Factors, Design Innovation and Safety Management courses. He holds a PhD in the field of Cognitive Ergonomics from the National Technical University of Athens - Greece, a graduate degree (D.E.A.) in Ergonomics, from the Conservatoire National des Arts et Métiers (C.N.A.M.) - France and an Engineering diploma in Production Engineering & Management science from the Technical University of Crete - Greece. His scientific interests span across all facets of the interaction of humans with technology with an emphasis on cognitive analysis and design. Before joining the academia, he has worked for several years in industry as a consultant in Ergonomics, Human – Computer Interaction and Occupational Safety, an activity that he continues practicing alongside his academic affairs. He is a certified European Ergonomist (EurErg) and currently serves as the President of the Hellenic Ergonomics Society and as national representative in the International Ergonomics Association. He has co-authored one handbook in Greek ("Introduction to Ergonomics").

Kostas Gkikas is a Product Designer, faculty member of the Ergonomics Lab of the National Technical University of Athens, and PhD candidate in the Department of Mechanical Engineering of NTUA. He is Certified European Ergonomist (EurErg) and is actively working as an Ergonomist consultant in Industry. He holds an Engineering Diploma in Products & Systems Design from the University of the Aegean - Greece with specialization in the field of Human Computer Interaction. He has worked as teaching faculty in the International Hellenic University and as a Graphic, Product and 3D Designer and Animator. His areas of interest include Cognitive Ergonomics, User-centered Design, UX Design, Design Thinking and Design Teaching.

Nicolas Marmaras is Professor of Ergonomics at the National Technical University of Athens, Dept. of Mechanical Engineering. He holds PhD in Ergonomics from the Conservatoire National des Arts et Métiers - France, a graduate degree (DEA) in Ergonomics from the Université Paris XIII - France and Diploma in Mechanical & Electrical Engineering -spec. Production Engineering, from the National Technical University of Athens - Greece. His research interests cover the areas of Cognitive Ergonomics, Ergonomic design, Human-Computer Interaction, User-centered design of IT systems, and Occupational Safety. He has undertaken and supervised many applied projects for Greek industries and organizations, dealing with improvement of working conditions, design or redesign of workplaces and products, plant layout, occupational safety and health, usability assessment of IT products, etc. He has authored three handbooks in Greek ("Introduction to Ergonomics", "Ergonomic Work Analysis" and "Occupational Safety & Health").

The present book is an introduction to Human Factors / Cognitive Ergonomics for interactive systems design, coupling applied theories of cognition with design methodology. The book came as a need to provide a comprehensive guide in user centered design destined primarily to students in information technology, engineering and design disciplines.

This work is the culmination of many years of experience in teaching Human Factors / Ergonomics together with relevant research and professional practice. Along with established and contemporary academic literature, this experience allowed the collection and development of a wide variety of examples and case studies of real-world systems from home appliances to large scale critical systems that the authors have found effective for transferring knowledge to students in a hands-on, experiential and dialectic manner.

The readers can either use the book as an introductory guide to human cognition, a methodological companion for the user centered design process or as a collection of examples and case studies to enrich their repertoire.

Textbook created in the context of the KALLIPOS+ project	
Funding by	Ministry of Education and Religious Affairs Public Investment Programme (ΠΔΕ), National Development Programme (ΕΠΑ) 2020-2025
Implimenting Body	ELKE, National Technical University of Athens (NTUA)
Operating Body	HEALink/NTUA Branch/Publications Unit
2nd phase duration	2020-2023
Aim	Creating open access electronic academic textbooks (more than 700) • Under- and Post-graduate textsbooks • Monographs • Translated open access textbooks • Reference Guides
Scientifically responsible	Nikolaos Mitrou, Professor, SECE, NTUA.
ISBN: 978-618-228-173-4	**DOI: http://dx.doi.org/10.57713/kallipos-406**

The textbook was funded by the Ministry of Education Public Investment Programme