Part I

A Long-Pondered Outfit

1 Extended Nature

> Instead of fruitless attempts to divide the world into things with and things without the essence of mind, or consciousness, we should examine the many detailed similarities and differences between systems.
>
> – Aaron Sloman,
> *The Structure of the Space of Possible Minds* (1984)

A PHENOMENAL DISPLAY is taking shape before our eyes. A variety of new behaviours present themselves embodied as computers, robots, enhanced humans, hybrids and collectives with sundry types of integration and communication. Can all possible behaviours, extant or extinct, actual or conceivable, natural or artificial, be embraced by the *machine kingdom*, the set of all computable interactive systems? What is the essence of each of these systems, how do they usually react and what are they able to do? The understanding and measurement of these behavioural features is not only a fascinating scientific challenge but an earnest need for society. It is needed for devising policies in all areas of life, from labour to leisure, and for the assessment of the effective progress and safety of the engineering disciplines behind this surge. What theory and tools do we have for this scrutiny? We only have some scattered pieces of the puzzle, and some of them do not match. It is time to integrate, sort and systematise these bits in the widest perspective. For this purpose, we must set a common conceptual ground upon which we can formulate old and new questions properly.

1.1 FACE THE DIVERSITY

Many biologists view our current time as a period of massive extinction. With the exception of some global catastrophes, this century will wash away more species than any other in history: millions, if not billions, of species.

The rate of this Anthropocene extinction is around 100 times the natural rate. Not yet tantamount in magnitude and diversity, but accelerating at a far greater rate, there is an opposite explosion of new creatures.

This massive explosion is about a different kind of breed. We call them computers, and they are all around us. They are constantly equipped with new types of communication and organisation, new bodies and interfaces and apparently whimsical ways of hybridisation. We have seen the rise of dustbots, digital pets, video game bots, robot swarms, online assistants, roboroaches, machine-animal herds, chatbots, machine translators, *animats*, algorithmic artists, crowdsourcing platforms and driverless cars. The contraptions are blending and pervading everything, leading to incipient cyborgs, enhanced – or atrophied – humans, human-assisted computers and emergent entities in social networks. Everyone can carry a chess master in her pocket device.

The physical capabilities and the external look of these artefacts are usually misleading about what they do and, most especially, about what they are able to do from a cognitive point of view. Endow a computer with facial expression and we will consider it more capable than what it really is. We will even empathise with it, as we do with dolls and puppets. In the opposite direction, however, one of the reasons behind the success of digital social networks, virtual worlds, online games and other *artificial ecosystems* is that they mostly rely on what their users do and say, on how they behave, on what they are capable of, and not on what they look like physically. This is a liberation experience for many people, who – perhaps for the first time in their lives – can be judged for what they really are.

The analysis of two interactive systems that differ on their physical capabilities and appearance is hampered by many confounding factors, if not simply thwarted by prejudices. Great effort has been put into the areas dealing with the evaluation of human behaviour to make testing procedures, such as exams, as independent as possible from physical traits and any other extraneous factors. Similarly, the evaluation of behavioural features in animals is performed with interfaces that try to isolate or discount all these confounding effects. Of course, this is not always easy or even possible, but the effort pays off when it is. Ultimately, the most elegant way of expressing the same idea was introduced by Alan Turing, with his famous imitation game, the Turing test: machines should be judged by what they do through a teletype communication.

We must then look at all this with the utmost neutrality, from an aseptic, unprejudiced standpoint. Panel 1.1 exemplifies the appropriate attitude when we observe a new organism.

When we make the effort of removing all the physical differences and look at each artefact or organism in a purely behavioural way, we get a much better

Panel 1.1
New to science

"There is a label on a cage that states simply, 'This machine is new to science'. Inside the cage there sits a small dustbot. It has bad temper. No bad-tempered dustbot has ever been found. Nothing is known about it. It has no name. For the mechanist it presents an immediate challenge. What has made it unique? How does it differ from the other dustbots already known and described?"

The preceding paragraph is adapted from Morris's 'The Naked Ape' (1967), where 'machine' replaces 'animal', 'dustbot' replaces 'squirrel', 'bad temper' replaces 'black feet' and 'mechanist' replaces 'zoologist'.

This paragraph represents the kind of unprejudiced standpoint about what the real *subject* of study is. Of course, this standpoint does not ensure a scientifically rigorous account, nor does it make the analysis any easier, but it sets a non-anthropocentric perspective.

understanding of what an organism truly is. Only with this perspective can we say that, for instance, a group of bacteria and a herd of sheep behave in a *social* way, despite the enormous physical differences between them and between their environments.

What are the organisms we need to scrutinise? The systems that are responsible for the new behaviours can be categorised into several groups:

- Computers: this refers to any type of computational behaviour, including any artefact that is designed with some kind of artificial intelligence (AI). We particularly emphasise those systems featuring machine learning, natural language processing, social interaction, complex perception and cognitive development (Russell and Norvig, 2009; Cangelosi and Schlesinger, 2015). AI systems that are responsive to situations they were not programmed for are becoming more versatile and relevant in our daily lives, doing or taking less mechanical, more cognitive jobs. Artificial general intelligence (AGI) is aiming at more ambitious goals such as open-domain question answering systems, developmental robotics and compositional learning.
- Cognitively enhanced organisms: here we refer to living organisms, such as "cyborg rats" with computer-controlled electrodes implanted in their brains (Yu et al., 2016), or more customary cyborgs, such as a deaf person with a cochlear implant. We also include humans whose cognitive abilities are altered by the use of any "tool of the mind" (Carr, 2011), such as a pen and paper, regarded in Plato's Phaedrus as "an elixir of memory and wisdom"

first, but a cause of atrophy afterwards. Actually, the notions of "extended mind" (Clark and Chalmers, 1998; Menary, 2010) or "natural-born cyborg" (Clark, 2004) make this concept very broad: every modern human is a cyborg. What we are envisaging is how humans can enhance their cognitive abilities by the use of technology (Cohen, 2013) or counteract age-related decay and mental disabilities in general through cognitive prosthesis (Hampson et al., 2012; Berger et al., 2012). For instance, how is our memory affected when we have access to the Internet? Are certain abilities being atrophied by the 'Google effect' (Sparrow et al., 2011)? Is technology making us 'stupid' and 'shallow' (Carr, 2008, 2011)?

- Biologically enhanced computers: there is no need for science fiction to see humans working for machines. This is already happening in several forms. Technically, 'human computation' (Von Ahn, 2005, 2009) "is simply computation that is carried out by humans" (Law and Von Ahn, 2011), seen as part of a more general problem a computer cannot solve efficiently. In practice, we see bots, some of them malicious, that rely on humans to recognise some difficult speech bits, to break authentication schemes (such as CAPTCHAs) or simply to be supervised. On one hand, this creates new questions about what cognitive abilities a computer can have with a bounded number of questions or interactions with a 'Human Oracle' (Shahaf and Amir, 2007). On the other hand, making a computer depend on people also creates availability and reliability problems. Humans make mistakes all the time.
- (Hybrid) collectives: any of the preceding groups can be structured in many different ways, including swarms or collectives combining humans, other animals and computers. A blind human with a guide dog is a traditional example. Actually, every hybrid can be seen as a collective, leading to new ways of co-operation, competition, communication and delegation. Video games and virtual worlds are playgrounds where the line between humans and computers is more blurred, and bots are evaluated by traits such as believability or enjoyability (Hingston, 2012). Online social networks are a good example of new types of interaction that might require some specific cognitive abilities or even change personality. Closely related, crowdsourcing is a paradigm whereby complex tasks are partitioned (by humans or computers) into smaller tasks that can be solved by humans or computers (Quinn and Bederson, 2011). A final integration stage can be a committee consensus (Kamar et al., 2012), but many other possibilities exist ("Crowds guiding AIs" and "AIs guiding crowds", Kittur et al., 2013).
- Minimal or rare cognition: many rare types of cognition are indeed new to science. Recent research in many different areas has been able to recognise new types of cognition in plants (Calvo-Garzón and Keijzer, 2011), bacteria (Van Duijn et al., 2006) and even forests at very different spatiotemporal

scales. Minimal cognition is also created and explored in virtual environments, with theoretical cognitive models (Beer and Williams, 2015) or artificial life organisms (Beer, 2015) that are endowed with minimal abilities to adapt to changes.

The emergence of so many new artefacts and systems requires a full reexamination of what behavioural features are and how they are measured. The first thing in this endeavour must be a more precise characterisation of the set of subjects to be analysed, beyond the traditional boundaries (humans, non-human animals and computers).

1.2 THE MACHINE KINGDOM

No matter how wild and diverse life may be, it is constrained by the rules of evolution and natural selection. Some behaviours are extremely unlikely, because the existence of organisms displaying them would require an improbable sequence of mutations and selections according to their odds of success and reproduction. The extant and even the extinct organisms are the "lucky ones", but one can consider "the set of all possible people allowed by our DNA" (Dawkins, 2000) or, still more generally, the set of all possible organisms (or genomes), the "Library of Mendel" (Dennett, 1995).

From a behavioural point of view, and before the discovery of DNA and the advent of computers, we have also been interested in systems that are not strictly living organisms: herds, social communities, symbiotic systems, etc., as well as mythical beasts, fictional characters and all "imaginary beings" (Borges, 1957). How can we characterise all behaviours displayed by all possible *interactive* systems? To answer this question we need to look at the principles of computation, as described in Panel 1.2.

There seems to be general agreement that the behaviour of a slug or a sponge can be simulated by a computer with arbitrary precision. By the evolutionary continuum, it is possible, in theory, to do the same with a mammal and, ultimately, with the human brain. The nuances appear when we discuss whether the resources and knowledge to do this will ever be available.

Under this view, all possible biological organisms and computers are machines, with extant ones being a subset. We can now give a definition of the whole set.

> **Keynote 1.1.** The **machine kingdom** is the set of all interactive systems taking inputs and producing outputs, possibly asynchronously, through interfaces, bodies, sensors and actuators, etc.

Panel 1.2
The physical Church-Turing thesis

The notion of a Turing machine, as defined by Alan Turing in 1936 (Turing, 1936, 1937), is still one of the most elegant models of computation. Turing showed that some problems are not decidable, i.e., there is no computable process that can answer them, such as the *halting problem*, the problem of determining whether an arbitrary Turing machine stops for an arbitrary input. Turing also introduced the notion of *universal* Turing machine (UTM), a Turing machine that can simulate any other Turing machine. Current computers are actually *resource-bounded* universal Turing machines, capable of *virtually* emulating other devices and implementing different programming languages, most of which are *universal*, also referred to as *Turing-complete*.

Alan Turing also postulated – and Alonzo Church in different terms – that any function that is effectively calculable is computable by a Turing machine. This is known as the Church-Turing thesis. One variant of the Church-Turing thesis, which goes beyond functions to interactive systems, is the *physical* Church-Turing thesis, stating that "every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means" (Deutsch, 1985, p. 99).

A mechanistic view of the human mind is a consequence of this thesis – every physical process in the universe, including those happening in every living thing, would be computable with finite resources (finite storage space and computational steps). We widen this view with machines that may be non-deterministic, such as probabilistic Turing machines, with a source of randomness (possibly through analog components or sensors), or non-functional, such as interactive Turing machines (or other computational models of interaction; Goldin et al., 2006), as many everyday computers.

Unless explicitly stated otherwise, we will restrict the attention to resource-bounded machines. The same algorithm running in a faster computer is, in terms of the machine kingdom, a different machine. We will generically refer to the elements of the machine kingdom as agents or subjects, especially when confronted with an environment, a task or a test. Figure 1.1 shows a simplistic Euler diagram where animals and humans are placed inside the machine kingdom.

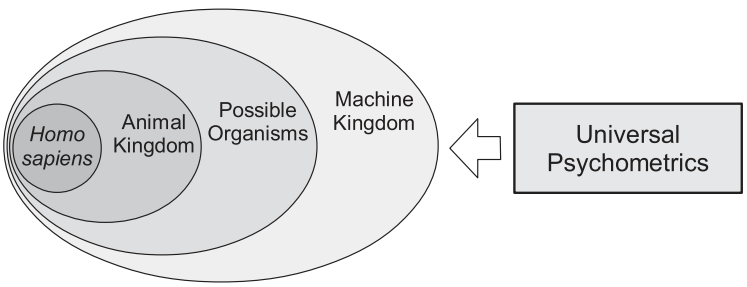

Figure 1.1. The machine kingdom embracing the animal kingdom. Though not explicitly shown on the Euler diagram, other living things, hybrids and collectives also belong to the machine kingdom.

There are several reasons why we make this move to a machinery scenario. The first reason is the avoidance of any kind of anthropocentrism or biocentrism. The second reason is that we want to make it more explicit that classical boundaries between the natural and artificial domains vanish. The third reason is that we want a more formal foundation for our analysis of subjects, their properties and abilities. For example, we can define distributions over the elements of the machine kingdom, which can take the role *populations* play for natural organisms.

The view represented by the machine kingdom is familiar in some disciplines. Artificial life is an area that bridges life produced in a virtual environment with the one produced in a physical environment. Artificial life goes beyond the constraints of the life on Earth (gaiacentrism) but also beyond the constraints of all organic life (biocentrism), by considering any possible organism. One significant feature of artificial life is that the environment is not considered as a separate entity from the organism. The identification of the subject in the environment and the channels of interaction are crucial. During most of this book, we will assume that the organism is well delimited. In Chapter 16, however, we will explore the notion of universal test, which requires the identification of the subject in the environment, and we will discuss the relevance of an appropriate interface.

Bio-inspired robotics is also an area halfway between the natural and the artificial, initially focusing on morphological and locomotive mechanisms but recently paying more attention to behaviours, closer to cognitive robotics and developmental robotics. In fact, some animal behaviours are emulated by artificial systems, known as *animats* (Wilson, 1991; Webb, 2009; Williams and Beer, 2010).

Cognitive science deals with the analysis of perception, attention, memory, learning, knowledge, emotion, reasoning and language in humans, other

animals and computer models. Basically, cognitive science studies the "Space of Possible Minds" (Sloman, 1984), with the term "mind" seemingly setting a preference for a very small, and not well defined, subset of the machine kingdom. In fact, the essence of mind was usually analysed at a more philosophical level with related but different concepts, such as materialism, determinism, free will, creativity, unpredictability and, ultimately, consciousness.

In general, the use of the term "mind" is now more comprehensive, as for Dennett's (1996), Goertzel's (2006) and Hall's (2007) "Kinds of Minds", Yudkowsky's "Mind Design Space" (2008) or Yampolskiy's "Universe of Minds" (2015a, p. 35), the latter also being used as a way of resuscitating the old (theistic) body-mind dualism (see, e.g., Carter, 2007, p. 12).

Following Sloman, we will not look for a qualitative essence of mind (or, more interestingly, person, whose characterisation will be seen in Chapter 17) but for a range of behavioural features that characterise all the elements in the machine kingdom. In fact, trying to avoid the use of the word 'mind' and its connotations, Sloman suggests the "space of possible 'behaving systems', to coin a neutral phrase" (Sloman, 1984), which is much closer, if not equal, to what we are referring to by the 'machine kingdom' here. Actually, some definitions of theoretical cognitive science just refer to "information processing systems" (Simon, 1980). In the end, our fixation on the behaviour of all interactive systems derives from our interest in the *measurement* of what systems do.

Once we have defined our *Cosmos*, in which humans are nothing more than a pale dot, what are we going to do with these *b*illions and *b*illions of machines? Our goal is to measure and classify them in terms of their behavioural features. This is what we call *universal psychometrics*.

> **Keynote 1.2. Universal psychometrics** is the analysis and development of measurement tools for the evaluation of *behavioural features* in the *machine kingdom*, including *cognitive abilities* and *personality traits*.

The use of the term *behavioural* feature instead of the more usual *psychological* or *mental* feature emphasises the general scope of the machine kingdom. Similarly to human psychometrics, universal psychometrics also covers attitudes, interests, beliefs and knowledge evaluation. In addition, while the term *cognitive development* is not explicitly included, universal psychometrics should also deal with the evolution of systems as a result of learning, education or other changes in the environment.

Do we have appropriate tools for the evaluation of these features for the diversity of systems in the machine kingdom? We can knock on several doors

here. Human psychometrics is the paramount discipline when we think of psychological measurement in general, and intelligence in particular. Much of what we know scientifically about cognitive abilities in humans originates from what psychometric research has done during more than a century. Countless test batteries are available for different groups and abilities. Intelligence quotient (IQ) tests are just one type of these tests, arguably the most popular and controversial. Can we use some of these psychometric tests for the variety of systems in the machine kingdom? This question will be addressed in more detail in Chapter 6, but we can anticipate that many psychometric tests are designed for some particular human populations, and their use in other populations is disputable.

Some of the new artefacts (e.g., animats) are closer to animals (or swarms) than humans. Comparative psychology is the discipline that has been concerned with the evaluation of the cognitive abilities of a range of species in the animal kingdom. Many hurdles had to be overcome, such as how to perform tests without the use of language, how to make animals focus on a task and how to choose the right interface and rewards. This encompassing effort gives comparative psychology more flexibility for the evaluation of all these new systems. We will reuse some of these ideas and techniques, but we will also see that many tests are focused on very specific animal traits and lack the breadth and depth of psychometric tests.

And what about artificial intelligence? Do we find tools in artificial intelligence to evaluate its artefacts? Turing's imitation game has deservedly become very popular, but it is not really the way artificial intelligence evaluates its systems. In practice, artificial intelligence uses specialised tests for each particular task. Chess-playing computers are evaluated with completely different tools than self-driving cars. Indeed, there is no general methodology for AI evaluation, with many different competitions and benchmarks being developed in the past decades. In the end, AI evaluation tools are task oriented rather than feature oriented.

The limited integration between these disciplines suggests a full overhaul of their principles and tools, much beyond a naive generalisation of human psychometrics. In particular, we must start with a re-examination of what features are to be measured.

1.3 THE SPACE OF BEHAVIOURAL FEATURES

Imagine for a moment there were an art to answer every solvable question, a procedure to unravel every intelligible mystery of the universe. Imagine you were given an *Ars Generalis*, the universal key that opens all locks.