UiO : Department of Informatics
University of Oslo

Reflections on Behavioural Computer Science

Christian Johansen , Tore Pedersen , Audun Jøsang

Research report 452, May 2016

ISBN 978-82-7368-417-2
ISSN 0806-3036

Abstract

The rapidly increasing pervasiveness and integration of computers in human and animal society calls for a broad discipline under which this development can be studied. We argue that to design and use technology one needs to develop and use models of humans/animals and machines in all their aspects, including cognitive and memory models, but also social influence and (artificial) emotions. We call this discipline Behavioural Computer Science, and propose that behaviour computer science models try to unify (models of) the behaviour of humans/animals and machines when designing any ICT systems. Incorporating empirical evidence for actual human behaviour instead of relying on assumptions about rational behaviour, is one of the contributions of this paper. We must also acknowledge that advancements in AI will give machines capabilities that from many perspectives are indistinguishable from those of humans/animals. We provide a few directions for approaching this challenge, focusing on modelling of human and machine behaviour as well as their interaction.

[0]*Address for Tore Pedersen:*
Center for Intelligence Studies, Norwegian Defence Intelligence School.
E-mail: tore.pedersen@feh.mil.no

Address for Audun Jøsang and Christian Johansen (neé Cristian Prisacariu):
Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo, Norway.
E-mail: {josang,cristi}@ifi.uio.no

The second and third authors were partially supported by the project Oslo Analytics funded by the IKTPLUSS program of the Norwegian Research Council.

This technical report is a long version of the paper [27].

1 Introduction

The marriage of ubiquitous computing and AI opens up for an environment where humans and animals will interact with autonomous systems that will be indistinguishable from humans/animals or systems directly controlled by humans. For simplicity we will in the following use the term 'human' in the implicit assumption that it also covers (intelligent) animals. Not only must humans relate to intelligent machines, the same machines must relate to humans and to other intelligent machines.

For humans, it could be a strange experience to interact with intelligent machines that might have contradicting traits of human behaviour. For example, a human normally reacts defensively or aggressively when physically attacked, but a human-looking robot might not react in this way if it is not programmed to do so. The implications of this disconnection between traits that are usually connected (i.e. human-looking and self-defence) is something humans will have to get used to, for better or for worse.

Our ethical compass should guide us to build intelligent machines that have desirable traits, whatever that might be. In order to achieve this goeal it is essential that we understand haw humans actually behave in interaction with intelligent machines, and this is a largely unexplored field. For example, what are the criteria for trusting an intelligent machine for which the intelligent behaviour *a priori* is unknown. Also, how can an intelligent machine trust humans with whom it interacts. Finally, how can intelligent machines trust each other. From a security point of view, the most serious vulnerabilities are no longer found in the systems but in the humans who operate the systems. In a sense, it is no longer a question of whether people can trust their systems, but whether systems can trust their human masters.

These are daunting challenges in the brave new world of intelligent ubiquitous computing and cyberphysical infrastructure. Three important fields of scientific study are fundamental to understanding and designing this infrastructure:

Behavioural Sciences working with systematic analysis and investigation of human behaviour through controlled and naturalistic observation and disciplined scientific experimentation. It attempts to accomplish legitimate, objective conclusions through rigorous formulations and observation. Examples of behavioural sciences include psychology, psychobiology, criminology and cognitive science.

In contrast to traditional, rational and normative approaches to how people should ideally behave (we use behaviour as a general concept that includes the subcategories judgement and decision making), behavioural sciences give scientific, empirical, evidence-based, and descriptive approaches

to how people actually make judgements and decisions. Thus, these two approaches are complementary: the rationalist describes the ideal behaviour, whereas the behaviouralist describes the actual behaviour.

Ubiquitous Computing and IoT is a new paradigm in software engineering and computer science where computing is made to appear anytime and everywhere. In contrast to desktop computing, ubiquitous computing can occur using any device, in any location, and in any format. A user interacts with the computer, which can exist in many different forms, including laptop computers, tablets and terminals in everyday objects such as a fridge or a pair of glasses. The underlying technologies to support ubiquitous computing include Internet, advanced middleware, operating system, mobile code, sensors, microprocessors, new I/O and user interfaces, networks, mobile protocols, location and positioning and new materials. The IoT is the connected aspect of ubiquitous computing.

Artificial Intelligence abbreviated AI, is the intelligence exhibited by machines or software. It is also the name of the academic field which studies how to create computers and computer software that are capable of intelligent behaviour. Major AI researchers and textbooks define this field as "the study and design of intelligent agents" [43], in which an intelligent agent is a system that perceives its environment and takes actions that maximise its chances of success according to some criteria.

We put these three areas under the umbrella called "*Behavioural Computer Science*" (abbreviated BCS). Any outcome of integrating these three areas would be called a BCS-*model*, which will always include the human aspects in some way or another. We would like to encourage research focus on the interactions between these three areas. This is illustrated in Figure 1. The intersections between any of two of these areas represent existing or new research disciplines.

The field of Interaction Design [46] and *Human-Computer Interaction* (HCI) studies how a technology product should be developed having the user in focus at all stages. Machine ethics is that part of the Ethics of Artificial Intelligence concerned with the moral behaviour of artificially intelligent beings. The field of Roboethics is concerned with the moral behaviour of humans as they design, construct, use and treat such beings. With the advent of ubiquitous computing, the Internet of Things (IoT) and advanced AI, the distinction and interface between computers and human becomes very blurred.

Artificial behaviour is an emerging discipline which focuses on understanding how intelligent systems behave from a macro-perspective, and not from a computer-program perspective. When intelligent systems become intelligent enough they will have psychological traits that can be studied.

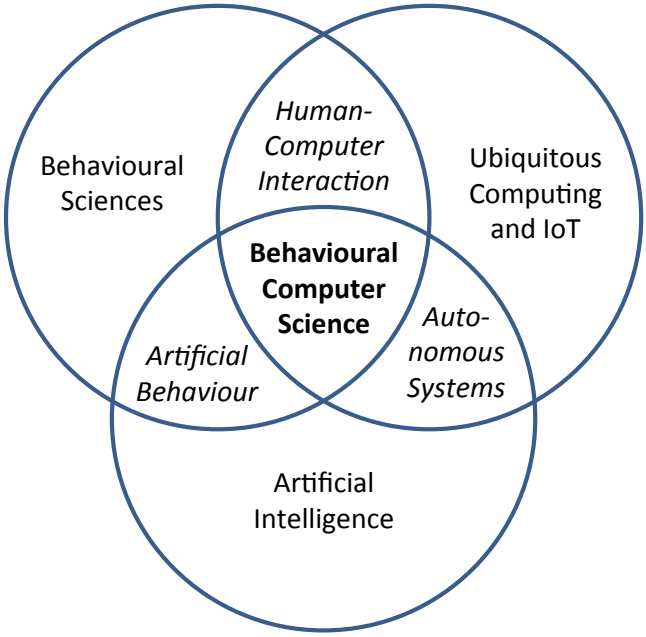

Figure 1: Conceptual Definition of Behavioural Computer Science

Models for how humans and intelligent machines interact can be understood in a general and inclusive manner, as any formally or mathematically grounded model useful in some way for building IT systems. We can think of probabilistic models, logical and formal models, programming and their many types of models, etc. The purpose of using models is to be able to tackle complexity; since we constantly see complexity becoming the norm for current day technology.

Computational trust becomes an aspect of machine learning or heuristics, that in turn will be part of IoT systems and other (semi-)autonomous controllers, or self-* systems. For such autonomous and powerful systems we need to study notions of trust [29], like trust of the user in the system, or of another interacting system or component.

To motivate modelling human and intelligent machine interaction we will use examples from security ceremonies and IoT, but our results could also be used in other situations, like to improve national intelligence systems.

2 Behavioural computer interaction

In domains where humans interact with technology it is necessary to understand human behaviour in order to capture or foresee possible actions taken by humans in interaction with the technology. We refer here to an understanding that can

be used by machines, thus through models that can be used in some forms of computations. If technology and their designers understand the typical tendencies of human cognition, emotion and action, it is easier for the resulting system to take into consideration how people actually behave, and adapt to this, instead of relying on assumptions about how they may behave.

As this implies, there are two primary approaches to including human models: one would follow the Rational Agent Model (e.g., [49, 50, 51]) and another the Behavioural Model of Human Agency (e.g., [54, 32]). The rationalist approach to explaining human behaviour is traditionally widespread in academia as well as in society in general. The rationalist tradition generally adheres to the view that people are rational agents that seek to maximize utility. Inherent in this approach lie the assumption that people know their ultimate goal, have the means to select the courses of action that are the most likely to lead to goal achievement, and the capability to carry out the appropriate courses of action. To do arrive at this end-state, people would need to have unlimited access to all information, the ability to discriminate relevant from non-relevant information, the cognitive capability to handle and analyze the interaction between the relevant informational components inherent in the possible courses of action, to calculate how the courses of action would lead to the possible end-states, and to foresee implications of the end-states. We shall soon see that these assumptions are seldom fulfilled, which leads us to focusing on the Behavioural Model of Human Agency.

One of the first proponents of the behavioural model was Herbert Simon, the 1978 Nobel Laureate in Economics. Simon found that people, when making real judgments and decisions, did not comply with the ideal that was assumed by the rationalist traditions. He was the first to coin the term Bounded Rationality [51] to describe the concept of non-ideal adherence to the rationalist assumptions and thus complemented the traditional rationalistic approach in the field of economics.

Prominent scholars in behavioural Science, after Herbert Simon, are Daniel Kahneman, the 2002 Nobel Laureate in Economics, and his late colleague Amos Tversky. Notable findings in Kahneman and Tversky's research [32, 23] is that people often rely on intuitive thinking when making judgments under conditions of uncertainty. Intuitive thinking, when employed inappropriately in conditions when instead analytic thinking would have been the correct cognitive strategy, often leads to biased – and consequently incorrect – judgments. Although the rationalistic approach is valuable in explaining how people should ideally make judgments under conditions of certainty, the behavioural approach is the better in explaining how people actually make judgments in uncertain conditions, and also identifying the bias-inducing psychological mechanisms people employ – mostly without conscious awareness [55].

Consider three examples where the behavioural approach to explaining human judgment has successfully enriched an existing academic discipline:

Behavioural Economics focusing on how people actually behave in economic contexts, as opposed to how they should ideally behave (e.g., [31, 32]), has been a fruitful addition to Economics;

Behavioural Game Theory focusing on how people actually behave in formal games, as opposed to how they should ideally behave (e.g., [14]), has enriched traditional Game Theory; and

Behavioural Transportation Research focusing on how people actually make choices in transportation and travel contexts, as opposed to how they are assumed to behave (e.g., [22, 41]), has been a fruitful addition to the traditionally rationalistic field of Transportation Research.

Our opinion is that Behavioural Computer Science can be one more fruitful collaboration between behavioural science and computer models, and this paper gives some venues of exploration. In particular, such collaborations could have a good influence on the field of Artificial Intelligence and Trust as well as Security.

Consider two examples of emerging fields which can be seen as part of BCS.

Security ceremonies have recently seen increased interest since they strive to involve the human aspect when designing and analysing security protocols [20, 44]. A few works have studied the human aspect of security breaches [60, 48, 2, 57]. Here an example is spear-fishing attacks, where we see technology developers taking the attitude that "the breach will occur", so they try to protect against it through e.g. network isolation of the infected system. We argue that cognitive models and models of social influence can give insights into how to build e-mail systems that can counter more effectively such targeted, well-crafted, malicious e-mails.

Home automation and ambient assisted living [3, 10] is one of the applications of IoT that is most closely interacting with humans occupying the house. Such systems need to learn patterns of behaviour, preferably distinguishing them among several occupants, adapt to temporary changes in behaviour, as well as interact and take control requests from the humans. A Nordic example can be smart heating system.

3 A reference model for BCS

To anchor our thoughts we will use a model introduced in [7], which we call *"the Bella-Coles-Kemp model"* and abbreviate as *BCK model*. This is a rather general and abstract model for any forms of human involvement in computer systems,

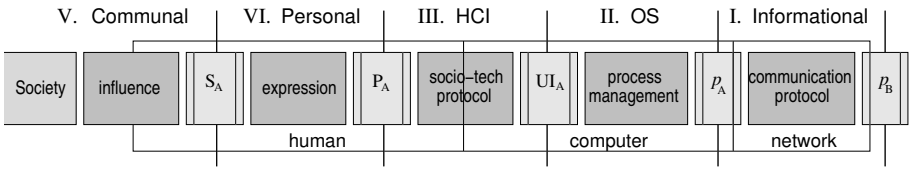

Figure 2: Bella-Coles-Kemp model (BCK model), taken from [7, Fig.1].

thus providing a good common basis for defining Behavioural Computer Science concepts.

The BCK model, pictured in Fig. 2, is intended to give abstractions and separations, still allowing for more details to be given for each of the vertical layers individually. We can see the behaviour sciences (like psychology and cognitive sciences) as a good ground for finding specific models for the layers V and IV; whereas for layer III we already have a good start in works from HAI. Traditional computer science studies layers I, II, and also partly layer III.

When explaining the BCK model it is good to make correlations with existing established concepts; and we choose here models for security protocols, usually based on the Dolev-Yao assumptions, and using specification languages like the applied pi-calculus [1]. Usually, security protocols are formed of the parties (or players) and the interaction medium they use for communication (or any other exchange of information). The parties are usually honest, whereas the intruder (attacker) controls the interaction medium. More than two parties can be involved in a protocol, but for our example purposes here let us consider only two honest parties, Alice and Bob. Third parties, usually dishonest, appear due to the ability of the intruder to disguise as a party in any number of protocol runs. The Dolev-Yao model defines the powers that the attacker has over the interaction medium, like power to delete, change, or insert messages, to and from any other party.

In the BCK model the parties form the light boxes, whereas the interaction medium forms the dark boxes. The parties appear at different layers of the BCK model and in different abstractions; i.e., the light boxes represent the players in the respective layer, which are abstractions of the parties or are controlled by the parties. In the layer I (also called "Informational") we encounter the processes p_A and p_B controlled by Alice respectively Bob, which are running the computers of Alice and Bob, communicating through the network, i.e., the dark box. Layer I would thus be studied by communication and networking researchers, and for security it could be subject to the standard Dolev-Yao assumptions. But this layer has also other aspects, like properties of the transmission medium (e.g., messages can be lost or not, or delayed and how much).

In BCK other players appear at the other layers: at layer II (also called "Operating System") *the user interface* UI_A associated to Alice, which interacts with

the computer process p_A, e.g., by sending information taken from the user required by the security protocol run by p_A, like a password or biometrics. The same UI_A interacts at layer III (also called "Human-Computer Interaction") with *a "persona"* P_A of Alice for some particular system. The persona has interaction at layer IV (also called "Personal") with *the self* S_A of Alice, which in turn is influenced by the Society through various social interaction protocols at layer V.

Players may interact only as part of a layer, and one layer may involve players pertaining to different users. Important to note is that in BCK one player usually is involved in two adjacent layers.

Research in computer science until now has mostly focused on layers I and II, and largely ignored layers III–V. We see layer II as pertaining also to the technological community, whereas layer V would pertain to the social sciences. Layer IV would be investigated more by the psychology researchers. Layer III on the other hand is at the interaction between technological and social sciences, with a rapidly evolving field, having terminology s.a.: HAI [9], user-centred design [8], interaction design [46].

One usefulness of BCK is to make explicit the need for collaboration between the fields of sociology, psychology, and technology, in order to tackle the complexities of current systems s.a. security ceremonies or IoT. One can very well focus on individual layers, but the BCK model brings the isolated results into the general picture which eventually needs to be handled in order to claim results and designs of practical use.

The BCK model is abstract and general, but we expect future research to detail all the new layers III–V, the same as has been done until now with the layers I and II. The interaction medium, the dark box, can be split into more fine-grained divisions, and each division would have its interaction protocol and assumptions. For layer II it is easy for computer scientists to bring their knowledge of operating systems design and see that a UI could consist of a screen and its driver, a display client like a browser displaying an input form, a keyboard with its drivers, and the many other components that transport the information between these many UI components and the end process p_A. But for the social protocols of layer V, completely different concepts and models should be used. One could study various means of social manipulation, and quantitative and qualitative measures could be devised for analysing their usefulness in terms of power to influence, e.g., depending on the social scale or training level of the users, i.e., the self S_A.

We have applied the BCK model to study security ceremonies in [26] where we focused on layer III and introduced probabilities as opposed to classic (rationalist view) non-deterministic models. The concepts discussed in this paper are concentrating on layers V and IV, and how these could be modelled and combined with the methods used in CS in layers I and II. Works on layer III constitute a good middle ground [37, 25, 15]. We particularly wish to focus our modelling

efforts for V and IV on behavioural research, and depart from the rational view and assumptions on human.

4 Behavioural and social aspects of humans and technology

In many domains, academic as well as professional and political, it is a generally held view that people make rational judgements and thus are assumed to think, act, decide and behave according to the rational agent model [51]. The rational agent model implies that people always strive to maximize utility.

With regard to the concept of *utility*, this is generally understood as the satisfaction people derive from the consumption of services and goods [39]. From this perspective, it is an overarching assumption that every individual knows his or her ultimate goals and also how to fulfil their goals. If one looks at utility from a psychology perspective, a problem arises because there is more than one definition of utility.

Experienced utility is the satisfaction one derives in the consumption moment, and thus the most valid measure of the general concept of utility.

But there are two other types of utility that are different from this:

Predicted utility (or, alternatively, *expected* utility) is the utility one predicts beforehand that one will experience in the future consumption moment.

Remembered utility is the utility one remembers having experienced in a consumption moment some time ago.

The problem with these three different aspects of utility is that the rational agent model implicitly assumes them to be equal, whereas empirical psychology research has found that these aspects reflect different utilities; the utility that one actually experiences may be different from both what one beforehand predicted and what one later will remember. The rational agent model does not take this into consideration when it generally regards predictions, experiences and memory of utility as representing the same type of utility.

Another problem with the rational agent model regards the concept of rationality. In this view it is assumed that people act strictly logical and rational in the pursuit of maximized utility. Inherent in this view, conditions are assumed to be certain, meaning that every individual is assumed to have unlimited access to all information and is also capable of analysing the relevant information needed to make a judgement, as well as calculating the outcome of every combination

of informational components, so that the best decision can be made. Of course, no single individual is able to adhere strictly to this model of rational behaviour, but the point here is that the rational agent model assumes rationality as a general principle, and does not concern itself with empirical evidence about actual behaviour (i.e., as opposed to inferred behaviour) as to whether this assumption is actually valid.

Although most proponents of the rational agent model realize that the model is more ideal than realistic, they nevertheless assume that individuals' errors of judgement or of informational processing (i.e., their failure to comply with the model and thus make perfect judgements), are non-systematic. Non-systematic errors mean that each single individual does not exhibit the same types of judgement errors as other individuals and, furthermore, these differ over time, i.e., errors made today are different than those made yesterday. In fact, it is acknowledged that people are not perfect judgement machines and thus make errors, but is assumed that mistakes are non-systematic and thus random [55, 31] (classically modelled through non-determinism).

To sum up, some errors in human behaviour often stem from the differences between predicted, experienced and remembered utility; e.g. when making judgements at time t_0 about some consumption related moment in the future at time t_1, one often disregards the fact that their current experiences will be different from their expectations. Errors may also occur as a consequence of making judgements in conditions under uncertainty, i.e., when the requirements of the rational agent model cannot be fulfilled.

Kahneman [31], and other behavioural scientists, questioned the explanatory powers of the rational agent model, because they could not make their empirical data fit the rational agent model. As psychologists – or behavioural scientists – they studied how people actually behave, as opposed to how they are assumed to behave according to the rational agent model. Thereby they provided empirical data that supported a new view – namely that people's judgement errors were not at all non-systematic and thus random, but in fact systematic; people tended to make the same kinds of misjudgmetnts as others did, and misjudgments made today are the same as those made yesterday. Thus, people's mistakes were more or less universal. Findings like these paved the way for a new model of human behaviour, namely the model of Bounded Rationality [51].

One major and universal finding in this new avenue of research is that there are two fundamentally different systems of cognitive processing [52, 31]:

System 1: Intuitive Thinking, is associative, effortless, emotion-influenced, automatic, and thus often operating without conscious awareness;

System 2: Analytic Thinking, is analytic, effortful, not influenced by emotions, sequential, controlled and thus operating with conscious awareness.

Because Intuitive Thinking is effortless and automatic, people have a tendency to rely heavily on this cognition mode in most everyday activities – where we automatically know how to judge, behave and decide – and it works fine. The problem is that we sometimes employ this automatic mode of thinking also in situations where we have less knowledge or experience. A failure to activate Analytic Thinking thus results in what is now commonly labelled as *biased judgements*.

Another major finding from behavioural sciences that is relevant to BCS is the discovery of four psychological mechanisms (also called *heuristics*) that are mostly responsible for the human tendency to make unwarranted swift judgements [23]. These four mechanisms – leading to biases in situations where we are uncertain – belong to Intuitive Thinking (which is thus sometimes called Heuristic Thinking). When we are making judgements under conditions of uncertainty, we are known to employ one or more of these heuristics, which often fail to make correct judgements. Let us now take a look at each of these heuristics and define their major characteristics.

The availability heuristic simply means that people make judgements based on what is easily retrievable from memory, or simply what comes easily to mind. Let us say that you are asked to list as many English words as possible that begin with the letter A. This is a simple task, because words beginning with the letter A are fairly easily retrievable from memory. Now, let us say that your friend is asked to list as many words as possible that have the letter A as the third letter in the word. This is a much more difficult task, because words with the letter A as the third letter are less easily retrievable. As a consequence of this, it may very well be that you will think that there are more English words beginning with the letter A, than words having the letter A as the third letter. If so, you have made an incorrect judgement caused by the availability heuristic.

The representativeness heuristic describes how people make a judgement based on how much the instance or the problem in front of them is perceived as similar to another known instance or problem. If the degree of perceived similarity is large enough, people will easily make incorrect judgements. Let us say that you are asked to give answers to questions under strict time constraints, and that you are asked to reply as fast as you can. In one of these questions you are shown a picture of a whale. If you incorrectly label this whale as a fish, you have made an incorrect judgement based on the representativeness heuristic.

The anchoring and adjustment heuristic implies that people – under conditions of uncertainty – without conscious awareness will establish an “anchor", and from this anchor adjust their judgement, often in the “right” direction,

although not to the point of accuracy. If you are in a condition of total uncertainty, even non-relevant information that you have either been primed with, or that is easily accessible from memory, can serve as an anchor.

The affect heuristic simply means that the current affective state may influence human judgements. For example, if you are in a positive mood, you may be more easily susceptible to deception and manipulation because of a tendency to making hasty and possibly incorrect judgements, whereas you may be less inclined to do so when you are in a negative mood.

To counteract the tendency towards the Intuitive Thinking, in order to make people less susceptible to the heuristics that may generate incorrect judgements, one possible intervention could be to "slow" people's actions down, inviting them to be consciously aware of their actions, and thereby make them employ System 2-thinking. The message that we get when trying to delete a file, saying "Are you sure you want to delete this file?" is an example of such an intervention.

For the spear fishing example, where one receives a malicious email from an address that resembles that of a known colleague. This is an attack that is difficult to counter because it *activates both the availability heuristic and the representative heuristic*; the user may or may not have no easily accessible information stored in the mind that may suggest that this is an hostile attack (susceptibility to the availability heuristic) and, furthermore, the user recognizes the email address as being from a near colleague (the representative heuristic). Additionally, when considering that malicious attackers could also employ mechanisms of social influence [16], such as the six principles of persuasion, wherein e.g., the concept of Authority (people have a tendency to obey instructions from authority figures) or Liking (people have a tendency to be more easily persuaded by people they like), they have access to a versatile tool-kit of psychological manipulations and deceptions, which they could use with malevolent intent. Thus, no alertness or caution is prompted.

Human choices and human prediction power are very important for interactions with computer systems, e.g. security can be influenced by poor predictions about the possibilities of attacks and attack surface can be wrongly diminished in the mind of the human, whereas wrong choices can incur safety problems. In [32] it is argued that it is difficult for a human to make accurate predictions about a situation or an experience (e.g., sentiment, preference, disposition) when the future forecasting time point t_0 is rather distant from the current time point t_0 on which the same experience is evaluated. The more distant this time point is, the more inaccurate the prediction (and thus the choice) will be.

5 Modelling for behavioural computer science

We are interested in how behavioural concepts could be mathematically modelled, and more importantly, how these behavioural models can be coupled and integrated with existing models from computer science. Thus, our study here pertains to the layers V and IV of BCK. We start discussing a very simple model, one similar to what we did in [26] for layer III, based on works from HCI [12, 18, 47] or from cognitive theories [36].

One point made by Kahneman and Thaler [32, 31] is that the circumstances (i.e., the context of the human and of the system) vary between the present t_0 and future t_1 time points. Four large areas of such *varying circumstances* can be identified:

The emotional state of the human, or the **motivational state** of the human might vary when t_0 and t_1 are distant.

The aspects of the choice, of the product, of the experience, that are considered as important or are made salient/observable at t_0, might not be present at t_1 or may be difficult to experience or observe at this later time point.

Memory of similar choices or experiences is important. If the memory is biased then the current choice and prediction for the future will be biased. Tests of memory manipulation have been made [30] and one observation is summarized as the *Peak/End Rule*, as opposed to the common belief that the monotonicity of the experience counts. Humans recall the experiences of the peak emotions or of the end of the episode.

Affective forecasting [41, 58] is a concept introduced to explain that when focusing on some aspect for making a decision, this aspect will inappropriately be perceived as more important at the time of (prediction and) decision than it normally will be at the time of experience.

We will work with a notion of "States" and changes between states (which we sometimes call "Transitions"). How exactly to model an *emotional or motivational state* is not trivial, and we discuss these in more details later. Let us focus now on *changes* between states. We have already discussed about *"temporal changes"*, i.e., changes that happen because of passage of time. These we can consider in two fashions:

gradual/continuous change in emotion or motivation happens over time (e.g., modelled with time derivatives, in the physics style); or

discrete changes where we jump suddenly from one value to a completely different value (e.g., think of motivation which can gradually decrease until it reaches a threshold where it is suddenly completely forgotten).

When we model *emotions* (as needed for *affective forecasting*, as well as for many aspect of the Self) we can start from the following concepts related to the *impact bias* [58]: the *strength* (or *intensity*) of an emotion and the *duration* [11]. Both of these can be quantified and included in a *quantified model of emotions*. Other temporal notions different than durations could be needed like *futures* or order *before/after*, for which there are well established models in computer science, e.g. temporal logics [34, 53, 4].

Another concept that we identify as influencing the Self is that of **events**, in the sense that *emotions are relative to events*. Events can be considered instantaneous and are sometimes modelled as *transitions* labelled by the respective event. The reason is that the event changes the state in some way, e.g., changes the memory of the Self, or attributes of various variables of the context as well as of the Self.

A cognitive explanation for people's biased retrieval of past experiences appears when we relate them to emotions. Whereas currently experienced emotions (related to a currently experienced event) are stored in the episodic memory, past experienced emotions (related to previously experienced events) are stored in the semantic memory. The semantic memory is largely susceptible to biases due to the influence of current beliefs about previously experienced emotions on the retrieval of memories [45]. Thus, memory may make people behave in ways unexplainable by the rational model.

These concepts contribute to defining *models for the predicted and the remembered utilities*, as well as how these models correlate with that of the experienced utility.

For *modelling a State* we can start by including the *aspects* that are of interest for the situation under study. Aspects could be modelled as logical variable that are true or false in some state, because they are either considered or not considered (i.e., observable/salient or not). The expressiveness of the logic to be used would be dependent on what aspects we are interested in; but we can start by working with predicate logic. Depending on the system being developed, we encourage to choose the most suited logic, e.g.: the SAL languages and tools which have been nicely used to describe the cognitive architecture of [48, Sec.2]; or one can use higher-order dynamic logic [24, Chap.3] and the tools around it like the KeY system [6].

Modelling memory and especially how can memory be manipulated and how the memory influences choices and thus transitions between states is not easy. Quite a few studies can be considered [17, 35], some of which are more close to models and to logics [13, 38, 56]. We can also use models from computer science

and logics like dynamic logic [24], used to talk about programming data structures, but also logics of knowledge and belief [21] which have a well developed models for how beliefs can be updated over time due to various changes [5].

Another important concept is that of *focusing illusion* [59], which is the illusion that an attribute/emotion that the human focuses on (since it is relevant for the respective emotion or activity or situation etc.) is more important than it actually is. The question is how to model the fact that an attribute is important? One alternative is to use *weights* and weighted models [19, 33]. We would then need empirical methods for automatically learning the weights as well as for measuring the importance/weight of the respective aspect. To fully capture focusing illusion we need to also include in the model a measure of how much *overrated* is the respective weight of the attribute in the current situation. Another question then is: How are these weights affecting the UI, persona, or the properties of the whole BCS-system?

These concepts cumulate in a model for the *Self*, involved in layers V and IV. Now the question is how does this model relate to the model of the Persona (that is involved in the HAI at layer III) and with the Society (at the outermost border of the system)?

The relation between the Self and the Persona can be seen as a simplification (called projection in more formal terms). The projection operation is done on a subset of the variables that make up the State of the Self, thus resulting in the state of the Persona. This projection would retain only those aspects that are relevant in the relevant context, i.e., in the context of the computer system being studied. This means that the projection operation should also be related to the model of the UI (i.e., the one between layers II and III).

But this simplification relation is not enough. We need to understand the interactions between the Self and the Persona. We can see **two interaction directions**:

from the Persona to the Self i.e., to the user with all the experiences, sensors, memory, thinking systems, heuristics, etc.; and

from the Self to the Persona i.e., to a simplified view of the user, specifically made for the UI and the system being studied.

Since a Persona is an abstraction of the human relevant for the interaction with a specific UI, then through the Persona we can see stimuli from the UI going to the Self, and influencing it. Therefore, the first communication direction can be seen as communications coming from the UI but *filtered through the Persona*.

For the second direction we see more the actions of *expression* (e.g., described by [48, 18]) that the Self makes out of the thoughts, reasoning, intuition, past experiences and memory models, into something relevant to this BCS-system and to the UI that the human interacts with. In consequence, we may say that the

Self interacting with the UI is filtered by the Persona we designated. But this Self is aware of more that just the UI: maybe she is aware of computer networking aspects (which pertain to layer I) or operating systems aspects of layer II (like how browsers work or how the operating system can be protected from bugs and viruses, whether an antivirus is installed or a firewall, etc.). All these examples are outside the direct visual interaction of the Persona with the UI, which is captured through the layer III.

Interactions at layers V and IV would be studied empirically through study of the Self and of Personas. A model would *start from general assumptions*, incorporated as prior information/probabilities. For a specific system, with a specific Persona defined, the model would need to be constantly updated by gradually learning from the empirical studies and evidences, thus *updating the priors*.

Because we use empirical evidences we need to introduce a notion of *uncertainty about the probabilities* that the studies reveal. Therefore, models of *subjective logic* [28] could be useful for expressing things like: "The level of uncertainty about this value given by this empirical study is the following.".

One would then be interested in applying standard analysis techniques like model-checking over these new models with uncertainty. This would allow to:

- Find ways how to protect the Self from malicious inputs and manipulation from the UI through the Persona.
- Find ways to protect the Self from the Social interaction in layer V, commonly called social-engineering attacks.

One type of such protective methods are known as *debiasing techniques* [42] which are useful for tackling focusing illusion. BCS would study how they could be integrated in the designed system, in the sense that the UI or the security protocol could implement features meant to manipulate the User in such a way that she would be prepared for a possible attack; or better, in such a way that they alert the user to the security aspects. Such features could involve: recollections, so that the same aspects of t_1 (now) are as in t_0 (the time point when the User has probably been trained to use the system).

6 Conclusion

We argued that concepts and findings from behavioural sciences can be translated into models useful for computer science. Such models could be used for analysing the BCS-systems using techniques such as automated model checking [4]. Moreover, behavioural models and related modelling languages can be used by system developers when making new BCS-systems to also consider the human interacting

with the system. We can already see promising results in this direction from using formal methods to analyse HAI systems [9] or human related security breaches [48].

As we have shown, there is now an abundance of research arguing that people behave and act in other ways than those assumed under the rationality paradigm. In consequence we proposed that computer science incorporates knowledge about actual behaviour in the design of systems that interact with humans. Psychology and Behaviour Sciences have by now provided a large amount of empirical evidence showing how human behavioral tendencies in many instances depart from strict rational assumptions and thus from what is inferred about behavior.

Psychology and knowledge about human behaviour and tendencies have often been employed by private interests for commercial purposes, for example with the aim of influencing or convincing people to purchase a specific product or service. Such knowledge has also been employed by political interests in order to convince people to endorse a particular political view. Even if some would argue that such approaches may have been employed in people's own interests, in order to have people make choices that are actually good for them, others would argue that the main purpose of such approaches is to serve the initiator – whether this is a private company or a political party.

Thus, some may fear that if, by following our proposal, models of human behaviour are made such that machines can work with them, then more easy it can be for a totalitarian regime to control people by using computers for mass-surveillance/-manipulation. One could even fear such models also in a western society because large corporations that control information, like those involved in search engines, social media, network corporations, or device and software producers, could be tempted to use such models in a negative manner, trying to gain control over their users. Indeed, any large corporation could possibly gain access to data from the previous providers and use behaviour models to corrupt their users, like one could imagine in industries s.a. tobacco, pharmaceutical, alcohol, or oil/gas.

Contrary to such uses of behavioural machine models, our main intention with the BCS approach is to take the perspective of the individual and to build HAI-interfaces that take into consideration human behavioural tendencies – in the interest of humans instead of commercial interests – with the aim of designing systems that empower individuals to make more correct judgements when interacting with the automated system.

Our proposal would serve the individual and thereby the society, as well as system designers and owners, due an increased knowledge about human behavioural tendencies. Additionally, when taking into consideration that society and all its functions rely on resilience both in infrastructure and in citizens, in order to maintain structure and daily life in all domains, it is necessary to avoid disruptions of

vital societal functions [40]. It is thus not an overstatement to claim that Behavioral Computer Science would also have valuable and positive implications for National Security (seeking to secure society) and for National Intelligence (seeking to identify threats to society).

References

[1] M. Abadi and C. Fournet. Mobile values, new names, and secure communication. In C. Hankin and D. Schmidt, editors, *POPL*, pages 104–115. ACM, 2001.

[2] A. Adams and M. A. Sasse. Users are not the enemy. *Com. ACM*, 42(12):40–46, 1999.

[3] J. C. Augusto, M. Huch, A. Kameas, J. Maitland, P. McCullagh, J. Roberts, A. Sixsmith, and R. Wichert, editors. *Handbook of Ambient Assisted Living*. IOS Press, 2012.

[4] C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.

[5] A. Baltag and L. S. Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[6] B. Beckert, R. Hähnle, and P. H. Schmitt, editors. *Verification of Object-Oriented Software: The KeY Approach*. Springer, 2007.

[7] G. Bella and L. Coles-Kemp. Layered Analysis of Security Ceremonies. In *Information Security and Privacy*, volume 376 of *IFIP AICT*, pages 273–286. Springer, 2012.

[8] N. Bevan. International standards for HCI and usability. *International Journal of Human-Computer Studies*, 55(4):533 – 552, 2001.

[9] M. Bolton, E. Bass, and R. Siminiceanu. Using formal verification to evaluate human-automation interaction: A review. *IEEE Trans. Sys., Man, and Cyb.*, 43(3):488–503, 2013.

[10] A. B. Brush, B. Lee, R. Mahajan, S. Agarwal, S. Saroiu, and C. Dixon. Home automation in the wild: Challenges and opportunities. In *SIGCHI*, pages 2115–2124. ACM, 2011.

[11] R. Buehler and C. McFarland. Intensity bias in affective forecasting: The role of temporal focus. *Personality and Social Psychology Bulletin*, 27(11):1480–1493, 2001.

[12] R. Butterworth, A. Blandford, and D. J. Duke. Demonstrating the Cognitive Plausibility of Interactive System Specifications. *Formal Asp. Comput.*, 12(4):237–259, 2000.

[13] M. D. Byrne and S. Bovair. A working memory model of a common procedural error. *Cognitive Science*, 21(1):31–61, 1997.

[14] C. F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

[15] J. M. Carroll, editor. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science.* Morgan Kaufmann, 2003.

[16] R. B. Cialdini. The science of persuasion. *Scientific American*, 2001.

[17] M. A. Conway, editor. *Cognitive Models of Memory*. MIT Press, 1997.

[18] P. Curzon, R. Rukšėnas, and A. Blandford. An approach to formal verification of human–computer interaction. *Formal Aspects of Computing*, 19(4):513–550, 2007.

[19] M. Droste, W. Kuich, and H. Vogler, editors. *Handbook of Weighted Automata*. Springer, 2009.

[20] C. Ellison. Ceremony design and analysis. Cryptology ePrint Archive Rep. 2007/399, 2007.

[21] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[22] T. Gärling, D. Ettema, and M. Friman, editors. *Handbook of Sustainable Travel*. Springer, 2014.

[23] T. Gilovich, D. Griffin, and D. Kahneman, editors. *Heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press, 2002.

[24] D. Harel, J. Tiuryn, and D. Kozen. *Dynamic Logic*. MIT Press, 2000.

[25] M. Harrison and H. Thimbleby, editors. *Formal Methods in Human-Computer Interaction*. Cambridge Univ. Press, 1990.

[26] C. Johansen and A. Jøsang. Probabilistic modeling of humans in security ceremonies. In A. Aldini, F. Martinelli, and N. Suri, editors, *Int. Workshop on Quantitative Aspects in Security Assurance (QASA)*, volume 8872 of *LNCS*, pages 277–292. Springer, 2014.

[27] C. Johansen, T. Pedersen, and A. Jøsang. Towards Behavioural Computer Science. In *10th IFIP WG 11.11 International Conference on Trust Management (IFIPTM)*, LNCS. Springer, 2016.

[28] A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212, 2001.

[29] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[30] D. Kahneman. Evaluation by moments, past and future. In D. Kahneman and A. Tversky, editors, *Choices, Values and Frames*, page 693. Cambridge University Press, 2000.

[31] D. Kahneman. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58:697–720, 2003.

[32] D. Kahneman and R. H. Thaler. Anomalies: Utility maximization and experienced utility. *The Journal of Economic Perspectives*, 20(1):221–234, 2006.

[33] M. Z. Kwiatkowska. Quantitative verification: models techniques and tools. In *Foundations of Software Engineering ACM SIGSOFT*, pages 449–458. ACM, 2007.

[34] Z. Manna and A. Pnueli. *The temporal logic of reactive and concurrent systems: Specification*. Springer-Verlag, 1992.

[35] I. Neath and A. M. Surprenant. *Human Memory: An Introduction to Research, Data, and Theory*. Thomson/Wadsworth, 2 edition, 2003.

[36] A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

[37] A. Newell and S. K. Card. The prospects for psychological science in human-computer interaction. *Human-Computer Interaction*, 1(3):209–242, 1985.

[38] K. Oberauer and R. Kliegl. A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55(4):601 – 626, 2006.

[39] R. L. Oliver. *Satisfaction: A Behavioral Perspective on the Consumer*. M.E. Sharpe, 2010.

[40] D. Omand. *Securing the State*. C Hurst Publishers, 2012.

[41] T. Pedersen, M. Friman, and P. Kristensson. Affective forecasting: Predicting and experiencing satisfaction with public transportation1. *J. Applied Soc. Psycho.*, 41(8):1926–1946, 2011.

[42] T. Pedersen, P. Kristensson, and M. Friman. Counteracting the focusing illusion: Effects of defocusing on car usersâĂŹ predicted satisfaction with public transport. *Journal of Environmental Psychology*, 32(1):30 – 36, 2012.

[43] D. Poole and A. Mackworth. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, 2010.

[44] K. Radke, C. Boyd, J. M. G. Nieto, and M. Brereton. Ceremony Analysis: Strengths and Weaknesses. In *Inform. Security and Privacy*, volume 354 of *IFIP AICT*, pages 104–115. Springer, 2011.

[45] M. D. Robinson and G. L. Clore. Episodic and semantic knowledge in emotional self-report. *Journal of Personality and Social Psychology*, 83(1):198–215, 2002.

[46] Y. Rogers, H. Sharp, and J. Preece. *Interaction Design: Beyond Human-Computer Interaction*. Wiley, 3rd edition, 2011.

[47] R. Ruksenas, J. Back, P. Curzon, and A. Blandford. Verification-guided modelling of salience and cognitive load. *Formal Asp. Comput.*, 21(6):541–569, 2009.

[48] R. Ruksenas, P. Curzon, and A. Blandford. Modelling and analysing cognitive causes of security breaches. *Innovations in Sys. Software Eng.*, 4(2):143–160, 2008.

[49] H. A. Simon. Rational decision making in business organizations. *American Economic Review*, 69(4):493–513, 1979.

[50] H. A. Simon. *Reason in Human Affairs*. Stanford University Press, 1983.

[51] H. A. Simon. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. MIT Press, 1997.

[52] S. A. Sloman. Two systems of reasoning. In Gilovich et al. [23], pages 379–396.

[53] C. Stirling. *Modal and Temporal properties of processes*. Springer-Verlag, 2001.

[54] R. H. Thaler and C. R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

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

[56] Y. Wang, editor. *Cognitive Informatics for Revealing Human Cognition: knowledge manipulations in natural intelligence*. IGI Global, 2013.

[57] R. West. The psychology of security. *Com. ACM*, 51(4):34–40, 2008.

[58] T. D. Wilson and D. T. Gilbert. Affective forecasting. volume 35 of *Advances in Experimental Social Psychology*, pages 345–411. Academic Press, 2003.

[59] T. D. Wilson, T. Wheatley, J. M. Meyers, D. T. Gilbert, and D. Axsom. Focalism: A source of durability bias in affective forecasting. *J. Person. and Social Psycho.*, 78(5):821–836, 2000.

[60] K.-P. Yee. User interaction design for secure systems. In *Information and Communications Security*, volume 2513 of *LNCS*, pages 278–290. Springer, 2002.