Towards Behavioural Computer Science

Christian Johansen, Tore Pedersen, Audun Jøsang

Towards Behavioural Computer Science*

Christian Johansen[1]**, Tore Pedersen[2]***, and Audun Jøsang[1]

[1] Department of Informatics, University of Oslo.
{cristi,josang}@ifi.uio.no
[2] Center for Intelligence Studies, Norwegian Defence Intelligence School.
tore.pedersen@feh.mil.no

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 (possibly artificial) emotions. We call this discipline Behavioural Computer Science (BCS), and propose that BCS models combine (models of) the behaviour of humans/animals with that of machines when designing ICT systems. Incorporating empirical evidence for actual human behaviour instead of relying on assumptions about rational behaviour is an important shift that we argue for. We provide a few directions for approaching this challenge, focusing on modelling of human behaviour when interacting with computer systems.

1 Introduction

The marriage of ubiquitous computing and AI opens up an environment where complex autonomous systems are heavily involved in the living and working environments of humans, often in a seamless fashion. Not only must humans relate to intelligent machines, but the same machines must relate to humans and to other intelligent machines.

Our ethical compass should guide us to build intelligent machines that have desirable traits, whatever that might be. In order to achieve this goal it is essential that we understand how humans actually behave in interactions with intelligent machines. 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.

* A long version of this paper is available as the technical report [13].

** The first author was partially supported by the project OffPAD with number E!8324 part of the Eurostars program funded by the EUREKA and European Community.
*** The second and third authors were partially supported by the project Oslo Analytics funded by the IKTPLUSS program of the Norwegian Research Council.

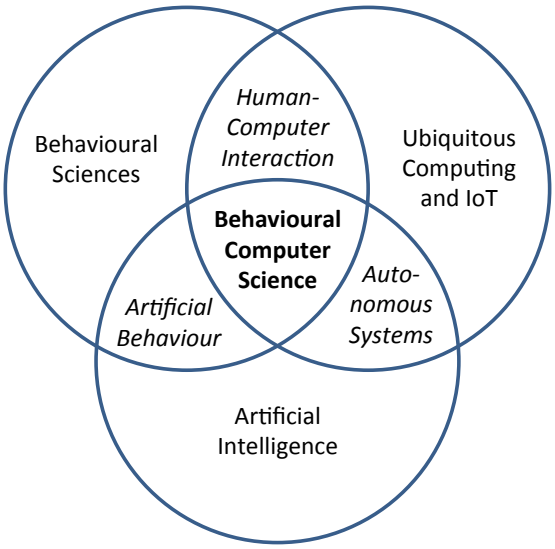

Fig. 1. Conceptual Definition of Behavioural Computer Science

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 giving scientific, empirical, evidence-based, and descriptive models for how people actually make judgements and decisions, as opposed to the traditional, rational and normative approaches that describe how people should ideally behave. Examples of behavioural sciences include psychology, psychobiology, criminology and cognitive science.

Ubiquitous Computing and IoT as the new paradigm in computer science where computing is made to appear everywhere, in various forms and everyday objects such as a fridge or a pair of glasses. Thus appear new forms of user interactions with such systems. The underlying technologies include Internet, advanced middlewares, sensors, microprocessors, new I/O and user interfaces. The IoT is the connected aspect of ubiquitous computing.

Artificial Intelligence studying how to create computers and computer software that are capable of intelligent behaviour. AI is defined in [23] as "the study and design of intelligent agents", 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 same umbrella called "*Behavioural Computer Science*" (abbreviated BCS) and illustrated in Figure 1. Any outcome of integrating models from these three areas would be called a BCS-*model*, which should also include aspects of human behaviour. We would like to encourage

research focus on the interactions between these three areas. The intersections between any two of these areas represent existing or new research disciplines.

Human-Computer Interaction (HCI) and more recently Interaction Design [24] studies how a technology product and its interface should be developed having the user in focus at all stages. With the advent of ubiquitous computing, the Internet of Things (IoT) and advanced AI, the distinction and interface between computers and humans becomes very blurred.

Models for how humans and intelligent machines interact can be understood in a general and inclusive manner, as any formally or mathematically grounded model used in building IT systems. We can think of probabilistic models, logical and formal models, programming and semantic models, etc. One purpose of using models is to understand and reduce complexity.

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 [15], like trust of the user in the system, or of another interacting system or component.

2 Behavioural aspects of humans and technology

When humans interact with technology it is necessary to understand human behaviour in order to capture or foresee possible actions taken by humans. We refer here to an understanding that can be used by machines, thus through models that can be used in forms of computations. If technology and its 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 in accordance.

Traditionally we find the Rational Agent Model (e.g., [27]) for explaining human behaviour, which generally adheres to the view that people are rational agents. However, it has been argued that the assumptions of the rational agent are seldom fulfilled, which leads us to focus on the Behavioural Model of Human Agency, as proposed by notable researchers like [28,18,10,31].

The rational agent model implies that people always strive to maximize utility, generally understood as the satisfaction people derive from the consumption of services and goods [20]. If one looks at utility from a psychology perspective, a problem arises because there is more than one definition of utility [18].

Experienced utility is the satisfaction derived in the consumption moment.

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 rational agent model implicitly assumes them to be equal, whereas empirical psychology research has found that these aspects reflect different utilities.

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 t_1, one often disregards the fact that their current experiences will be different from their expectations.

Rationality assumes that people act strictly logical, in the pursuit of maximized utility. In consequence, conditions are assumed to be certain, with humans having unlimited access to all information and also capable of analysing the relevant information needed to make a judgement, as well as calculating the outcome of every combination of informational components. However, behavioural scientists [18] questioned the explanatory powers of the rational agent model, because they could not make their empirical data fit the rational agent model. A new view, supported by empirical data emerged, showing that people's judgement errors were not at all 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. Moreover, *human errors* appear also as a consequence of making judgements in conditions under uncertainty, i.e., when the requirements of the rational agent model cannot be fulfilled.

One universal finding in this new avenue of research is that there are two fundamentally different systems of cognitive processing [29,17]:

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. The problem is when this automatic mode of thinking is applied in situations where we do not have enough knowledge or experience. A failure to activate Analytic Thinking in these situations may lead to systematic errors, also labelled **biased judgements**.

Another finding from behavioural sciences that is relevant to BCS is four psychological mechanisms (also called *heuristics*) that are mostly responsible for the human tendency to make unwarranted swift judgements [10]. These belong to Intuitive Thinking and lead to biases in situations where we are uncertain.

The availability heuristic explains how people make judgements based on what is easily retrievable from memory, or simply what comes easily to mind.

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.

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 explains how the current affective state may influence human judgements, e.g., when in a positive mood, one may be more easily susceptible to deception and manipulation.

To counteract the tendency towards the Intuitive Thinking, one possible intervention is to "slow" people's actions down, thereby making 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.

A spear phishing attack, where one receives a malicious email from an address that resembles that of a known colleague, is difficult to counter because it *activates both the availability heuristic and the representative heuristic*; the user may not have easily accessible information stored in her 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).

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 [18] 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_1 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.

3 Modelling for behavioural computer science

We anchor our thoughts using concepts from a model introduced in [3], which we call *"the Bella-Coles-Kemp model"* and abbreviate as *BCK model*. More details not necessarily relevant for this section can be found in [13, Sec.3] or in [12] where we used the BCK model in the context of security ceremonies. We will call the human *the Self*, which can be *influenced* by the *Society*, e.g., through social-engineering methods. The Self is *expressed* for a particular computer system as a *Persona*, understood as a collection of attributes relevant for a particular system interaction. The Persona is *interacting* with the system through the *User Interface (UI)*, often called *socio-technical protocol*. Socio-technical protocols have been studied in the Human Computer Interaction and related fields [5,4,24].

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. We discus a few aspects, some related to works from HCI [7,25] and from cognitive theories [19].

Kahneman and Thaler [18,17] argued 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 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 [16] 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 [21,33] – the process of predicting future emotions – explains how when focusing on some aspect for making a decision, this aspect may 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 call "Transitions"). Modelling an *emotional or motivational state* is not trivial, so let us look at the *changes* between states first. 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 style of physics); 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).

For modelling *emotions* (as needed for *affective forecasting* and many aspect of the Self) we start from the two concepts related to the *impact bias* [33]: the *strength* (or *intensity*) of an emotion and the *duration*. Both 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 [30,2].

Also influencing the Self are **events**, since *emotions are relative to events*. Events can be considered instantaneous and modelled as *transitions* labelled by the event name, because an event changes the state in some way, e.g., changes the memory of the Self, or attributes of the context as well as of the Self.

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

For *modelling a State* we start by including the *aspects* 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 [26, Sec.2]; or one can use higher-order dynamic logic [11, Chap.3] for more complex structures.

The relation between the Self and the Persona can be seen as a simplification (or projection). 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 respective 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.

Besides the simplification relation 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*.

The second direction considers actions of *expression* (e.g., described by [26,7]) that the Self makes out of the thoughts, reasoning, intuition, past experiences and memory models, filtered by the Persona and directed towards the UI.

Such interactions would be studied empirically, looking at the Self and Personas. A model *starts from general assumptions*, incorporated as prior probabilities. For a specific system, with a specific Persona defined, the model would constantly be updated by learning from the empirical studies and evidences.

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* [14] 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:

- Identify ways to protect the Self from malicious inputs and manipulation from the UI through the Persona.
- Identify ways to protect the Self from social-engineering attacks.

One type of protective methods are *debiasing techniques* [22], useful for countering biases caused by the focusing illusion. A BCS system could implement, part of the UI or the security protocol, features meant to manipulate the User in such a way that she would be prepared for a possible attack. 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 in using the system).

4 Further work

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 [2]. 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 [5] or human related security breaches [26].

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., [17,18]), 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., [6]), 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., [9,21]), 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 models and computer models.

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

Security ceremonies propose to involve the human aspect when designing and analysing security protocols [8]. A few works have studied the human aspect of security breaches [26,32]. An example is phishing e-mails where 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.

Ambient assisted living [1] is one application of IoT that is most closely interacting with humans. Such systems need to learn patterns of behaviour, distinguishing them among several occupants, adapt to temporary changes in behaviour, as well as interact and take control requests from the humans.

References

1. Augusto, J., Huch, M., Kameas, A., Maitland, J., McCullagh, P., Roberts, J., Sixsmith, A., Wichert, R. (eds.): Handbook of Ambient Assisted Living. IOS Press (2012)
2. Baier, C., Katoen, J.P.: Principles of Model Checking. MIT Press (2008)
3. Bella, G., Coles-Kemp, L.: Layered Analysis of Security Ceremonies. In: Information Security and Privacy. IFIP AICT, vol. 376, pp. 273–286. Springer (2012)
4. Bevan, N.: International standards for HCI and usability. International Journal of Human-Computer Studies 55(4), 533 – 552 (2001)
5. Bolton, M., Bass, E., Siminiceanu, R.: Using formal verification to evaluate human-automation interaction. IEEE Trans. Sys., Man, and Cyb. 43(3), 488–503 (2013)
6. Camerer, C.F.: Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press (2003)

7. Curzon, P., Rukšėnas, R., Blandford, A.: An approach to formal verification of human–computer interaction. Form. Aspects Comput. 19(4), 513–550 (2007)
8. Ellison, C.: Ceremony design and analysis. Cryptology ePrint Archive Rep. 2007/399 (2007)
9. Gärling, T., Ettema, D., Friman, M. (eds.): Handbook of Sustainable Travel. Springer (2014)
10. Gilovich, T., Griffin, D., Kahneman, D. (eds.): Heuristics and biases: The psychology of intuitive judgment. Cambridge University Press (2002)
11. Harel, D., Tiuryn, J., Kozen, D.: Dynamic Logic. MIT Press (2000)
12. Johansen, C., Jøsang, A.: Probabilistic modeling of humans in security ceremonies. In: QASA. LNCS, vol. 8872, pp. 277–292. Springer (2014)
13. Johansen, C., Pedersen, T., Jøsang, A.: Reflections on Behavioural Computer Science. Tech. Rep. 452, Department of Informatics, University of Oslo (April 2016)
14. Jøsang, A.: A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–212 (2001)
15. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems 43(2), 618–644 (2007)
16. Kahneman, D.: Evaluation by moments, past and future. In: Choices, Values and Frames, p. 693. Cambridge University Press (2000)
17. Kahneman, D.: A perspective on judgment and choice: Mapping bounded rationality. American Psychologist 58, 697–720 (2003)
18. Kahneman, D., Thaler, R.H.: Anomalies: Utility maximization and experienced utility. The Journal of Economic Perspectives 20(1), 221–234 (2006)
19. Newell, A.: Unified Theories of Cognition. Harvard University Press (1990)
20. Oliver, R.L.: Satisfaction: A Behavioral Perspective on Consumer. Sharpe (2010)
21. Pedersen, T., Friman, M., Kristensson, P.: Affective forecasting: Predicting and experiencing satisfaction with public transportation. Journal of Applied Social Psychology 41(8), 1926–1946 (2011)
22. Pedersen, T., Kristensson, P., Friman, M.: Counteracting the focusing illusion: Effects of defocusing on car users predicted satisfaction with public transport. Journal of Environmental Psychology 32(1), 30 – 36 (2012)
23. Poole, D., Mackworth, A.: Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press (2010)
24. Rogers, Y., Sharp, H., Preece, J.: Interaction Design: Beyond Human-Computer Interaction. Wiley, 3rd edn. (2011)
25. Ruksenas, R., Back, J., Curzon, P., Blandford, A.: Verification-guided modelling of salience and cognitive load. Formal Asp. Comput. 21(6), 541–569 (2009)
26. Ruksenas, R., Curzon, P., Blandford, A.: Modelling and analysing cognitive causes of security breaches. Innovations in Sys. Software Eng. 4(2), 143–160 (2008)
27. Simon, H.A.: Reason in Human Affairs. Stanford University Press (1983)
28. Simon, H.A.: Models of Bounded Rationality: Empirically Grounded Economic Reason. MIT Press (1997)
29. Sloman, S.A.: Two systems of reasoning. In: Gilovich et al. [10], pp. 379–396
30. Stirling, C.: Modal and Temporal properties of processes. Springer-Verlag (2001)
31. Thaler, R.H., Sunstein, C.R.: Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press (2008)
32. West, R.: The psychology of security. Com. ACM 51(4), 34–40 (2008)
33. Wilson, T.D., Gilbert, D.T.: Affective forecasting. Advances in Experimental Social Psychology, vol. 35, pp. 345–411. Academic Press (2003)