University of Warsaw

Faculty of History

Institute of Information and Book Studies

Jan Kaczmarek

Experience representation in information systems

PhD thesis prepared under the supervision of
prof. dr hab. inż. Mieczysław Muraszkiewicz

June 1, 2013

Summary

This thesis looks into the ways subjective dimension of experience could be represented in artificial, non-biological systems, in particular information systems.

The pivotal assumption is that experience as opposed to mainstream thinking in information science is not equal to knowledge, so that experience is a broader term which encapsulates both knowledge and subjective, affective component of experience, which so far has not been properly embraced by knowledge representation theories. This is the consequence of dominance of behaviourism and later cognitivism in the XXth-century science, which tended to reduce mind and experience respectively to behavioural expressions and discrete states relating mindful creature to external world, meanwhile the processes of knowing to manipulations with symbols.

We support the view that traditional knowledge representation approaches will not suffice to embrace the entirety of mental phenomena. We propose that in order to understand, represent and model the thinking and behavioural processes of mindful entities in information systems we need to look into the phenomenon of experience beyond the boundaries of knowledge. At the same time we propose to look at experience in a more structured way and try to capture it in formal terms, making it amenable to symbolic representation, being aware at the same time of innate limitations of symbolic representations compared to the natural representations in biological bodies.

Under the paradigm of mind intentionality, which assumes that minds have this special intrinsic feature that they can relate to external word and thus are *about* external world, it can be asserted that experience is one in all intentional mind state composed of knowledge that is the intentional contents of this state, the world-to-mind relation, meanwhile its inseparable subjective component is composed of subjective feelings of the mindful individual corresponding to this intentional mind states. If so, we propose that experience can be defined as two-dimensional mental phenomena consisting of mental states that have both knowledge and affective component. Consequently we suggest that experience can be represented as pairs of elements of sets K, and A, where K represents knowledge, hence contents of remembered intentional states of mind (i.e. intentional contents of experience), whereas A represents affect, i.e. the subjective qualitative component of experience.

Importantly, it does not particularly matter if we define experience as a set of mind states or a mind state process for assessing if the overall relation between knowledge and subjective experience that we have outlined above is valid. Whether there is *knowing* rather then *knowledge* or *experiencing* rather than *experience* which seems increasingly a contemporary principle, remains a fascinating philosophical, ontological to be more specific, question, however it falls beyond the scope of the thesis and therefore we shall not concentrate on it herewith.

Furthermore we propose that the subjective component of experience is also intrinsically intentionalistic, but meanwhile the intentionality in case of knowing is directed outward, to the external world, in case of feeling it is directed inwards to the within of the experiencing mindbody. We tap into the contemporary thinking in the philosophy of mind that the primordial, intrinsic intentionalistic capacity of mind is non-linguistic, as there must be other more primordial, non-linguistic form of intentionality that allows human children, as well as other language-capable animals, to learn language in first place. Contemporary cognitive neuroscience suggest that this capacity is tightly related to *affect*. We also embrace the theories of *consciousness* and *self* coming from brain scientists such as Damasio and Panksepp who believe that there is a primordial component of self, a so called protoself composed of the raw feelings coming from within the body, which are representations of bodily states in the mind, and have strictly subjective character. Therefore we can look at this compound of primordial feelings as a mirror in which external world reflects via the interface of the senses. This results in experience that has this conceptually dual, yet united within the conscious mindbody, composition of intentional contents that is knowledge and subjective component that is built up by feelings coming from within the experiencing mindbody. For it is problematic to state sharply either that this composition is dual or united we can refer to these two separately considered aspects of experience either as *components* or *dimensions*.

In this thesis we pay particular attention to the role the affective component of experience plays in the behaviour of organisms, and we use the concept of rational agency to discuss the relations between agent experience and behaviour. This role is primarily about motivation and experience vividness, i.e. how easily experiential states can be retrieved from memory. The affective dimension of experience determines the drivers for agent action and influences the remembering and forgetting (memory) processes that experience is prone to. We reflect on how the above presented framework could enhance one of the most popular rational agency models: the Believes Desires Intentions model (BDI) based on Bratmann's account of practical reason that has dominated information science and artificial intelligence literature.

Inspired by Davidson, who opposing Hume's account that the passions (desires) drive action while reason (belief) merely directs its force, concluded that

> "(...) belief and desire seem equally to be causal conditions of action. But (...) desire is more basic in that if we know enough about a person's desires, we can work out what he believes, while the reverse does not hold." (Davidson, 2004)

we conclude that in so far as BDI model approaches them, desires are sort of beliefs. Indeed a desire in the above sense is a *verbalised desire*, i.e. in order for a proposition to be included in the deliberation an agent must have internally verbalize it and accept it by which he converts it into a belief. As a result an agent acquires a belief about its desire.

Apart from desires made thus explicit and becoming beliefs there are implicit experiential states that directly influence behaviour, these are not embraced by the Desires set in the BDI and other instrumentalist rationality models as these currently do not have adequate forms of representation. If this is so, the BDI models looses its D creating a gap which must be filled in, which we try to do with the subjective dimension of experience. Under such an account each belief, either the proper one or about the desire, represented formally with a proposition should have an extra component added which would stand for the subjective affective state to this belief. Some preliminary suggestions how this could be implemented are proposed and discussed.

The central proposition of this thesis states that experience, broadly understood as the entirety of contents and quality of a conscious mind state, can be satisfactorily represented in information systems, and any information system which objective is to emulate natural agent behaviour with satisfactory faithfulness cannot do without a sound experience representation framework. To achieve this it is necessary to realize and accept, based on convincing evidence from neuroscience, that the missing subjective component of experience is *affect* that forms and integral part of natural agent's experience, and determines, or at least impacts profoundly the behaviour of natural agents. Relating affect to knowledge would result in a satisfactory approximation of experience. It is to realize as well that the subjective dimension of experience, classified as affect, is not entirely private, subjective epiphenomenal entity but rather can be studied in objective terms as neurological correlates in the brain following account of emotion and affect as fostered by contemporary neuroscience. By identifying affective correlates of intentional contents of states of mind, which build up knowledge, we can exploit a broader concept *experience* for the purpose of more accurate emulation of natural agents' thinking process and behaviour in information systems.

This thesis presents and discusses a bulk of evidence coming mainly from three fields: information science, philosophy of mind and cognitive neuroscience that led us to the above stated conclusions, as well as establishes a framework for experience representation in information systems.

Streszczenie

Tytuł pracy: Reprezentacja doświadczenia w systemach informacyjnych.

Przedmiotem pracy są sposoby reprezentacji subiektywnego wymiaru doświadczania w niebiologicznych systemach sztucznych, w szczególności w systemach informacyjnych.[1]

Głównym założeniem wyjściowym pracy jest to, że doświadczenie, w przeciwieństwie do dominującego w nauce o informacji podejścia, nie jest tożsame z wiedzą, doświadczenie jest ogólniejszym pojęciem, które obejmuje zarówno wiedzę jak i subiektywny, afektywny składnik doświadczenia, co do tej pory nie został prawidłowo ujęte przez teorie reprezentacji wiedzy. Działo się tak w ostatnich kilku dekadach głównie za sprawą dominacji redukcyjnego behawioryzmu, a później kognitywizmu w nauce XX wieku, które to prądy sprowadzały umysł i jego doświadczanie odpowiednio albo do wypadkowej obserwowanych zachowań albo do relacji pomiędzy podmiotem poznania a środowiskiem zewnętrznym, a procesy myślowe do manipulacji symbolami.

Postulujemy stanowisko, że tradycyjne metody reprezentacji wiedzy nie są wystarczające, aby ująć całokształt procesów myślowych i stanów mentalnych właściwych istotom obdarzonym umysłem. By zrozumieć, reprezentować i modelować w systemach sztucznych, takich jak systemy informacyjne, procesy myślowe i zachowanie organizmów należy przyjrzeć się doświadczaniu w szerszym ujęciu, wykraczającym poza granice wiedzy, obejmującym całość treści poznania z jego aspektami obiektywnymi jak i subiektywnymi. Jednocześnie proponujemy przyjrzeć się doświadczeniu w bardziej uporządkowany sposób pozwalający ująć je w formalne ramy, co umożliwiłoby próbę systematycznej reprezentacji doświadczenie za pomocą symboli, będąc

[1] Termin doświadczenie (ang. *experience*) używany jest w całej pracy jako termin obejmujący całość treści poznania, zarówno jego obiektywny jak i subiektywny wymiar. Jest to ważne dla zrozumienia celu i przedmiotu pracy, ponieważ w języku polskim termin ten częściej stosowany jest w innych choć bliskich znaczeniach, w tym w odniesieniu do eksperymentu, wiedzy empirycznej zdobytej w wyniku przeprowadzania doświadczeń (eksperymentów), wiedzy opartej na eksperymencie (interakcji z otaczającą rzeczywistością) czyli *wiedzą empiryczną* lub *wiedzą a posteriori*, czasem w znacznie zawężającym znaczeniu wrażeń czy doznań zmysłowych, albo umiejętności, wiedzy praktycznej (jak wykonywać zadania) związanej z wieloletnim praktykowaniem danej aktywności czy specjalizacją. Pozornie lepszym odpowiednikiem angielskiego *experience* jawi się staropolskie słowo *eksperiencja*, jednak ono częściej stosowane jest w nawiązaniu do eksperymentowania, co nie odpowiada znaczeniu które nas interesuje.

przy tym w pełni świadomym immanentnych ograniczeń reprezentacji symbolicznych systemów formalnych w porównaniu z naturalnymi reprezentacjami stanów myślowych w materii biologicznej organizmów żywych.

Podpierając się teorią intencjonalności umysłu, paradygmatem w filozofii umysłu zakładającym, że umysł ma tę szczególną fundamentalną właściwość, że może odnosić się do zewnętrznego świata, a więc że może być *o* lub być *skierowany do* (ang. *about* lub *directed at*) przedmiotów i stanów rzeczy w świecie, można stwierdzić, że doświadczenie w ujęciu ogólnym jest stanem intencjonalnym złożonym zarówno z wiedzy, która stanowi intencjonalną treść tego stanu, tj. relację umysł-świat oraz z nieodłącznego komponentu subiektywnego, odpowiadającego subiektywnym odczuciom towarzyszącym danym intencjonalnym stanom umysłu. Jeśli tak, to doświadczenie można postrzegać jako niedualny fenomen dwuwymiarowy składający się ze stanów mentalnych, które obejmują zarówno wiedzę jak i nieodłączny składnik afektywny. W związku z tym proponujemy, że doświadczenie może być reprezentowane jako nierozłączne pary elementów zbiorów K i A, gdzie K oznacza wiedzę, czyli zawartość zapamiętanych intencjonalnych stanów umysłu (tj. intencjonalna treść doświadczenia), natomiast zbiór A odpowiada afektowi, stanowiącemu subiektywny, jakościowy wymiar doświadczenia.

Istotnie, nie ma przy tym znaczenia czy postrzegamy doświadczenie jako zjawisko dyskretne, zbiór stanów umysłu czy ciągły proces umysłowy, by ocenić zasadność zaproponowanego przez nas powiązania w doświadczeniu wiedzy z subiektywnymi stanami umysłu. Niezależnie od przyjętej postawy ontologicznej wobec kategorii wiedzy, czy przyjmiemy za właściwsze odniesienie się do niej jako do ciągłej czynności odpowiadającej czasownikowi "wiem" (ang. *knowing*) czy raczej jako uniwersalium odpowiadającego dyskretnemu stanowi, kwantowi wiedzy (ang. *knowledge*), a w przypadku doświadczenia czy będziemy mówić o procesie doświadczania (ang. *experiencing*) zamiast o doświadczeniu w kategoriach uniwersalnych (ang. *experience*), nasze ustalenia pozostają w mocy.[2]

Ponadto postulujemy, że subiektywny składnik doświadczenia ma również własności intencjonalne, tym niemniej podczas gdy intencjonalność składowej doświadczenia odpowiadającej wiedzy jest skierowana na zewnątrz, ku zewnętrznej wobec podmiotu poznania rzeczywistości, intencjonalność subiektywnych uczuć skierowana jest do wewnątrz doświadczającego podmiotu. Posiłkujemy się przy tym współczesnymi teoriami umysłu, świadomości i psychologii rozwojowej, uznającymi, że pierwotna, wrodzona intencjonalna właściwość umysłu ma niejęzykowe podłoże, jako że musi istnieć niezależna bardziej pierwotna od językowej forma intencjonalności pozwala-

[2]Zarysowane tu rozróżnienie pozostaje fascynującym filozoficznym problemem, pozostającym w ścisłym związku z zakresem tematycznym niniejszej pracy jednak nie mieszczącym się w jej limitach objętościowych, w związku z czym ten wątek nie może zostać podjęty w dalszych częściach wywodu, poprzestańmy zatem jedynie na obserwacji, że z dwóch powyższych postaw ta pierwsza jawi się jako współcześnie dominująca.

jąca rozwijającym się organizmom, dzieciom w przypadku człowieka, jak również innym naczelnym rozwijającym umiejętności komunikacji językowej, nauczyć się języka w pierwszej kolejności. Współczesna neuronauka poznawcza utrzymuje, że ta bardziej pierwotna forma jest ściśle związana z afektem, subiektywnymi doznaniami i emocjami. Ponandto, wspieramy się współczesnymi teoriami *świadomości* i *ja* o rodowodzie neuronaukowym, takich autorów jak Damasio i Panksepp, którzy uznają, że istnieje pierwotny składnik ja, tzw. *proto-ja* (ang. *protoself* złożone z czystych odczuć pochodzących z głębi doświadczającego ciała, które reprezentowane są przez umysł jako stany całego biologicznego organizmu, mają ściśle subiektywny charakter. Stąd też, możemy spojrzeć na umysłowy obraz tych pierwotnych, czystych uczuć jak na zwierciadło w którym odbija się świat zewnętrzny za pośrednictwem zmysłów, z czego wynika koncepcyjnie dualny, jednak ujednolicony w świadomości obiektu poznania charakter intencjonalnych stanów umysłu złożonych z treści intencjonalnej oraz subiektywnych stanów uczuciowych pochodzących z wnętrza doświadczającego organizmu. Ponieważ trudno jest rozstrzygać ostatecznie co do dualnej czy jednolitej natury doświadczenia, głównie z powodu ograniczeń języka i historii współczesnej nauki, która zdominowana jest przez wpływy dualizmu kartezjańskiego, w całej pracy omowiając powyższe dwa aspekty doświadczenia posługujemy się określeniami *składniki* wymiennie *komponenty* lub *wymiary* doświadczenia.

W niniejszej pracy zwracamy szczególną uwagę na afektywny wymiar doświadczenia i rolę jaką odgrywa w celowym działaniu podmiotu poznania. Odwołujemy się w tym celu do koncepcji racjonalnego, celowego działania i filozofii praktyczności (racjonalnego rozumu) jak również ekonomicznej teorii wolnego wyboru aby prześledzić związek doświadczenia w ujęciu przez nas proponowanym z zachowaniem racjonalnych podmiotów (ang. *rational agents*). Rola ta przede wszystkim obejmuje kształtowanie motywacji oraz żywotność doświadczenia, tj. stopień jego utrwalenia w pamięci długotrwałej, z której doświadczenie może być przywoływane do pamięci krótkotrwałej w momencie świadomego wnioskowania na temat planowanych działań lub automatycznego wykonywania zachowań autonomicznych. Afektywny wymiar doświadczenia kształtuje działania podmiotu racjonalnego oraz wpływa na procesy pamięciowe, zapamiętywanie i zapominanie którym doświadczenie się poddaje.

W pracy podjęte zostały również rozważania na temat możliwości zastosowania zaproponowanych ram teoretycznych reprezentacji doświadczenia w najbardziej popularnym i powszechnie stosowanym w nauce o informacji, teorii i technikach sztucznej inteligencji modelu i formalnym systemie opisującym racjonalność, tzw. modelu BDI (ang. *Believes, Desires, Intentions*) opartym na logice praktyczności zaproponowanej przez Bratman'a.

Z przemyśleń innego filozofa praktycznego rozumu Davidsona, który w opozycji do Humowskiej tezy stwierdzającej, że żądze (ang. *passions*) decydują o pragnieniach (ang. *desires*) napędzają działania, podczas gdy rozum (przekonania) jedynie kierują siłami tego oddziaływania, skonkludował, iż:

> „(...) przekonania i pragnienia wydają się jednakowymi przyczynami działania. Jednakże (...) pragnienia jawią się bardziej podstawowymi tak, że jeśli wiemy wystarczająco dużo o czyichś pragnieniach możemy wnioskować co do jego przekonań, jednakże odwrotność tego stwierdzenia nie zachodzi." [3]

wywodzimy, że w odniesieniu do sposobu w jaki model racjonalności BDI je ujmuje, pragnienia są szczególną postacią przekonań. W istocie, pragnienie w tym ujęciu jest *zwerbalizowanym pragnieniem*, tj. aby pewien sąd mógł być uwzględniony w procesie rozważania (ang. *deliberation*) co do działania, podmiot racjonalny musi go zwerbalizować i zinternalizować, czyli włączyć do zbioru swoich przekonań. W konsekwencji podmiot tworzy przekonanie co do pragnienia, co mogłoby z kolei skutkować zredukowaniem pragnień, których bezpośredniość oddziaływania na zachowanie podmiotu nie jest inaczej realizowana w modelu. Jednakże, oprócz pragnień w ten sposób ujawnionych i włączonych do zbioru przekonań istnieją ukryte stany doświadczenia które bezpośrednio wpływają na zachowanie z pominięciem procesu internalizacji, czy niezależnie od niego. Tego typu pragnienia nie znajdują odzwierciedlenia w zbiorze pragnień modelu BDI ani mu pokrewnych, ponieważ nie istnieją odpowiednie formy ich reprezentacji. Jeśli tak, to system BDI zawiera istotną lukę którą należy uzupełnić, czego my próbujemy dokonać posiłkując się subiektywnym komponentem doświadczenia. W takim podejściu każde przekonanie, zarówno to właściwe jak i to będące uświadomionym pragnieniem, ujęte formalnie w postaci sądu logicznego powinno być doposażone atrybutem odpowiadającym subiektywnemu, afektywnemu stanowi odczuwanemu w momencie pozyskania lub przywołania do pamięci operacyjnej stanu umysłowego odpowiadającego danemu przekonaniu. Wstępne propozycje co do sposobu implementacji takiego rozwiązania zostały zaproponowane i omówione w niniejszej pracy.

Teza niniejszej rozprawy stwierdza, że doświadczenie, rozumiane szeroko jako całość treści i jakościowej wartości poznania złożonej z intencjonalnych stanów umysłu, może być w sposób zadowalający reprezentowana w systemach informacyjnych, oraz że dowolny system informacyjny, którego celem jest emulowanie zachowań podmiotów racjonalnych przy zadowalającym poziomie wiarygodności, musi uwzględniać subiektywny, jakościowy wymiar doświadczenia. W tym celu niezbędne jest przyjęcie, na podstawie przekonywujących wyników badań na gruncie neuronauki, że subiektywny, jakościowy komponent doświadczenia kształtowany jest przede wszystkim przez *afekt*, stanowiący integralną część doświadczenia podmiotu poznania, oraz czynnik determinujący zachowania podmiotów obdarzonych umysłem. Odwzorowanie wiedzy w zbiór możliwych stanów afektywnych skutkuje zadowalającą aproksymacją doświadczenia, innymi słowy szacowanie doświadczenia poprzez kojarzenie danego kwantu wiedzy z towarzyszącym mu uczuciem subiektywnym daje zadowalające rezultaty. Tym sa-

[3] Tłumaczenie własne autora na podstawie oryginału (Davidson, 2004)

mym proponujemy także, że subiektywny wymiar doświadczenia, nie jest całkowicie prywatnym, ukrytym przed obserwatorem ulotnym fenomenem, przeciwnie, może on być aproksymowany z obserwacji zachowania, relacji z introspekcji oraz bezpośrednio, obiektywnie studiowany w oparciu o neurologiczne korelaty w systemie nerwowym, głównie mózgu, zgodnie z aktualną praktyką współczesnej neuronauki. Rozpoznając afektywne neurologiczne korelaty intencjonalnych treści umysłu, które składają się na wiedzę, możemy korzystać z nowego zasobu terminologicznego i konceptualnego jakim jest *doświadczenie* celem opracowania sprawniejszych sposobów emulacji zachowania i procesów myślowych podmiotów racjonalnych z wykorzystaniem systemów informacyjnych.

W niniejszej pracy dokonujemy gruntownego przeglądu i omówienia najnowszych wyników badań naukowych, korzystając głównie ze źródłowych tekstów angielskojęzycznych, z kilku zazębiających się w obrębie zadanego tematu dyscyplin naukowych: nauki o informacji, filozofii umysłu oraz neuronauki, które zainspirowały nas do sformułowania wyżej streszczonych wniosków oraz stworzyły punkt wyjścia dla zaproponowanej przez nas struktury reprezentacji doświadczenia w systemach informacyjnych.

Contents

List of Figures

Preface

Like supposedly many other young people charmed by science I have been developing my scientific interests being driven by a curiosity about a question of a fundamental sort – why people, animals, any other living creatures feel like doing anything? The question could be made even more general – why electrons feel like spinning? – but let us ration ourselves to the living organisms. With this kind of general question ahead one can take up a wide variety of specific research paths across many disciplines from theology to quantum physics. I started with economics where human (consumer) behaviour and choice (decision making) were some of the central foci of interest and dug further across philosophy and cognitive psychology, further to cognitive neuroscience to end up in information science with two main conclusions:

- The decision making process is a product of a neurological information processing system which defines each and every decision made by any agent in the biosphere.

- A relatively neglected aspect in the study of this information processing systems, especially in the area of artificial systems that undertake to emulate natural systems, is its affective and subjective, first-personal dimension. Putting it in more plain yet a bit oversimplified terms the emotional dimension of human experience (consciousness) in which knowledge and information is generated, stored and processed has not received enough attention from scientists dealing with theory of information.

This thesis is inspired and driven by my deep believe that after a century of domination of rationalism in the positivist sense we will witness a period in science and society when much more due attention will be paid to the affective and the unconscious side of the human and animal nature, and biological life at large, which is the basic for understanding who we the human beings are. As the domination of positivist thinking in XX century catalysed the flourishing of philosophy of language, normative economics, behavioural psychology, development of information technology, the turn to the emotional and inner dimensions of biological and mental phenomena should let to the better understanding of non-linguistic forms of intentionality of living creatures, human and animal consciousness, human decision making and

problem solving, increased understanding of social behaviour and will open the way to many practical applications that will make the world a much more liveable place.

Recently, in a journalistic essay, I risked a purposively controversial claim that we seem to be living in increasingly romantic times. As the boundary between mental and biological, rational and emotional in the traditional sense of the terms is blurring, breaking the dualistic Cartesian conceptual framework, the emotional, more primitive, evolutionarily older, shared with other animals, side of human nature has gained more recognition and interest from scientists and engineers. The turn to the emotional, the biological (the animal) and subjective (spiritual to certain extent) is bringing about associations with romanticist thought. *Turn* may not be the ideal word in this case however, as it implies backlash from the realm of reason and objectivity meanwhile it is rather to suggest that the acceptance on an equal footing of both narratives: (i) the rational, objective, third personal on one hand and (ii) the emotional, subjective, first-personal on the other is anticipated.

The phenomenon apparently is entering not only the mainstream intellectual discourse but also has become a part of mass culture, to which a contemporary essay by Zygmunt Bauman (Bauman, 2011) testifies. Bauman observes that

> "*"Spirituality"* is the recently most recommended and earnestly desired value."[4]

Although Bauman makes this observation in a different context, warning against instrumentation and commercialization of morality as *lapis philosophorum* of pan-market urging for yet another, this time possibly limitless, exploitation field for consumerism, in his essay he depicts contemporary social reality in which inner, subjective mental states gain on significance in domains which have been traditionally conceived as rational, objective, such as: economy, science and politics.

Undoubtedly, the above mentioned developments have been prompted, to some extent, by the fascinating results from human brain studies, advanced by modern brain imaging technologies and earlier by studies on animal models, unravelling that the affective processes in the brain provide the very basis for human and animal behaviour and are merely mediated by higher cognitive processes such as conscious judgement and reasoning in particular. This tells us that the human mind seen as an information processing machine is in fact operating with information that is "affected", in other words information is coloured by subjective feelings experienced by the minded self. There is virtually no information encoded in biological systems that is fully cleared from this affective "stain". Meanwhile artificial systems, information systems in particular, are all the contrary, which creates a gap between these two distinct realms, the result of which are various complications and limitations of applications of artificial systems to solving real-world problems.

[4] Author's own translation from original text in Polish. Inverted commas by Bauman.

I believe this thesis is a humble contribution to the development of an emerging paradigm in information studies, the *affective theory of information* (ATI). This paradigm recognises that the mainstream traditional theories of information should be extended in a way that they embrace the subjective, affective dimension of human experience. Information and knowledge that can be encapsulated with language cannot be represented completely without this missing part: the subjective feelings of a minded self that receives, stores and processes information, as they have impact, have meaning, make the difference.

This thesis proposes that the affective dimension of experience can and should be represented in artificial systems so that these could be better applied to solving real-world problems that are intrinsically complex. I wish it makes a tiny step forward by providing theoretical grounds and justification, by putting together relevant scientific evidence from various disciplines, for defining models of human experience representation and modelling in artificial systems.

This representation and modelling methods and techniques will potentially open the way for many practical applications across many disciplines, for instance: (i) they can help construct information systems that are more usable and ergonomic for the user, (ii) they can allow for the construction of more believable artificial agents both in entertainment applications like computer games and virtual and mixed-reality worlds, as well as business and social applications that require emulation of human behaviour. Moreover, (iii) it can contribute to the construction of more usable human-machine interfaces and interaction paradigms as well as the development of more accurate decision-support systems. In this way the thesis will contribute apart from the field of information theory to such fields as: affective computing (computer science), customer experience management (marketing and management science), agent-based macroeconomic simulations (economics), complex problem solving (applied computer science), human-machine interaction and artificial intelligence (computer science).

This thesis could not have been completed without kind support of many people. In particular I would like to thank: Professor Mieczysław Muraszkiewicz for constructive feedback as this thesis progressed, Dr Dominik Ryżko from Warsaw University of Technology for join scientific endeavour in the area of experience modelling and the benefits brought by his sharp mind along the way. I shall also thank my friends: Dr. Jakub Lebuda, Łukasz Iwasiński and Michał Wendorff for participating in my personal development and years of challenging debates. I owe my special acknowledgements to Dr. Jakub Lebuda whose support allowed me to complete this project.

I dedicate this thesis to my sons Stanisław and Józef.

Chapter 1

Introduction

The purpose of this introductory chapter is to present thesis objectives, rationale and background. In this chapter we will put the thesis subject matter in the interdisciplinary context of the literature study that has been performed during the elaboration of this thesis as well as we will provide arguments for why the subject we address is valid and timely. It will end with a short overview of the thesis structure and contents presented chapter by chapter.

1.1 Thesis objectives

The overall goal of this thesis is to work out theoretical framework for building methods of human experience representation in information systems supporting emulation of natural agents behaviour. This overall goal confronts several challenges for which reason the following specific objectives has been defined within the overall goal:

1. To investigate and untangle the nature of experience, map its relationships with knowledge and purposeful behaviour.

2. To track the differences between experience representation and knowledge representation and show the significance of these differences with regard to information system efficiency.

3. To devise new ways for an improved representation of agent's affective dimension of experience and propose a conceptual framework for embracing affect as integral part of agent's experience.

4. To set up the theoretical framework that would link affect and knowledge with natural agent's purposive behaviour resulting from conscious decision making.

5. To work out a comprehensive mental phenomena representation framework that would capture the natural agent experience to the largest possible extent, and

would map agent's experience, including knowledge and affect with voluntary action.

6. To work out an experience representation and processing model that would suit the purpose of information systems that support emulation of natural agent behaviour.

1.2 Thesis background and rationale

This thesis reports the results of author's literature research and theoretical considerations on the nature of human experience and its representation in artificial systems, information systems in particular, which embraced several, traditionally separate, but today tightly interrelated fields: information science, artificial intelligence, philosophy of mind, cognitive psychology, neuro science, and economics. Importantly, these endeavours were inspired by several recent developments within this fields that revolutionized the way human experience is perceived and mapped onto practical reason. These developments will be introduced below together with a summary of main conclusions that are fundamental for this thesis. For purpose of clarity and consistency this introduction has been structured around the following horizontal topics:

1. Advancements in and limitations of information theories, information processing technologies and information systems,

2. Affect and its role in understanding, defining and modelling rationality.

3. Methodological inspirations: Leibnitz's *Calculemus!* and Wilber's *orienting generalizations*.

1.2.1 Advancements in and limitations of information theories, information processing technologies and information systems

Information theory and information processing technologies, fast development of which started after Claude Shannon published his ground breaking paper (Shannon, 2001) laying down the theoretical basis for digital information storage and communication, closely relate to the subject matter of this thesis: the representation methods of knowledge and experience. Claude Shannon's work has not only started the industrial digital revolution but also impacted profoundly philosophy, providing grounds for epistemological view later labelled computer functionalism, according to which human brain to human mind is like computer hardware to computer software, so that brain is literally and merely an information processing machine and all contents of mind is information. This by default eliminates subjective qualitative

dimension of experience from the central focus of philosophical discourse. Following some contemporary philosophers we believe that such a reductionist account of conscious mind is incomplete. Meanwhile information systems suffer clearly from this reductionist, functionalistic legacy, which imposes the most severe limitations upon them, we believe there may be ways for representation of subjective, affective components of experience with properly adapted knowledge representation methods. The era of dominance of computer functionalism has been slowly yet progressively coming to end end over past decades to which neuroscience largely contributed. All in one, this thesis has been elaborated during very exceptional times for information, computer science and the study of human mind. On the one hand these are the times of unquestioned triumph of information science, on the other hand the discipline is clearly approaching a critical moment in its history, which may either be a major breakthrough or a beginning of its steady decline. Affective computing inspired by affective neuroscience and computational theories of mind inspired by cognitive neuroscience and philosophical non-reductive physicalism are setting the scene for the years to come.

The purpose of this section is to outline the historical and current developments within information science that on the one hand render the representation of natural agents' experience in artificial systems possible at all, at least to a certain extent, but on the other to which current important limitations and bottlenecks of information systems and information processing technologies owe they origins. These developments, to be considered from both a scientific as well as socio-economic perspective along the current section have importantly influenced the line of thought of this thesis. Before we get to this part however it seems beneficial to better locate the subject matter of the thesis within the information science at large.

As the overall goal of this thesis is to work out a framework for developing methods of human experience representation in experts systems supporting emulation of purposive behaviour of natural agents under different decision situations, the subject matter of the thesis relates to the information science in a variety of ways, where the most important contact points are: (i) information systems (ii) knowledge representation methods and (iii) expert systems, the three branches embraced by the domain of information science. On top of that interdisciplinary branches such as computational neuroscience, robotics or human-computer interaction (HCI) clearly show that information theory and information systems play an important role in disentangling the processes that shape human, and other natural agent, purposive behaviour.

Information system is a tool used for data collection and storage as well as answering user queries and reporting [1]. The thesis proposes that human experience could be represented in information systems the same way knowledge is. Being

[1] See Chapter 2 for terminological discussion

able to map experience onto information and knowledge stored and processed by information systems could open the way to many new applications of these systems, such as emulation of natural agents behaviour, decision support systems, multi-agent simulation sistems, modelling social processes, intelligent system engineering, to name a few. Importantly the sole capability to store and emulate user experience in information systems has an intrinsic value. We can consider an information system that is capable of collecting, storing and processing information about experience of a particular type of user or natural agent and answer questions such as: 'what is the current experience of user x related to object y?', 'what user experience attributed to user x had resulted from the interaction with object y by user x?', etc.. Such a system could be applied to improved human machine inferfaces, web interfaces for instance, e-commerce applications, customer relationship management information systems, and many more.

Modern information systems encompass other additional functions of which knowledge dissemination and discovery are most relevant. The ultimate goal is that information systems are capable to provide intelligible and meaningful answers to more sophisticated questions rather then simply returning results of data quarries. In case of experience emulation system this is about questions such as: 'when should an event y be triggered to improve experience of user x'; 'What is the state of experience of user x'; 'When user x is in the moment of peak experience?'; 'What the event y should be like to improve user's y experience of object z', 'How presentation of object v or participation in event y will alter the experience of user x', etc. These present basic challenges to an interdisciplinary discipline of artificial intelligence (AI).

As indicated by Muraszkiewicz (Muraszkiewicz, 2011) the 'meeting point' of artificial intelligence and information systems is knowledge representation ($\mathrm{AI} \cap \mathrm{IS} = \mathrm{KR}$). The methods and techniques of knowledge representation are the central focus of artificial intelligence research as the possibility of representing knowledge in the form of symbols enables its collection, storage and processing by machines, computers in particular. There is a tight feedback loop that binds AI with other disciplines apart from information science: philosophy (e.g. study of consciousness, cognition, reasoning, ethics), biology and medicine (bio-physiological aspects of human cognition, reasoning, information processing by neural system, brain in particular) mathematics and logics (e.g. reasoning, argumentation, mathematical modelling of cognitive and decision processes); psychology (cognitive psychology in particular), computer science (e.g. intelligent systems, machine learning, expert systems, algorithms) and virtually any other discipline at the level of application, for instance: management, finance and banking, robotics, machine engineering, physics, chemistry and many others. This is important as it shows the practical value of the topics undertaken by this thesis, to be more specific the application of experience representation methods in systems that emulate natural, mostly human, agents' voluntary action, or systems that undertake

to create believable virtual characters in applications such as computer games and other audiovisual productions, as well as human-computer interaction technologies.

The relevance of KR to the thesis objectives has two facets: (i) representation of experience in information systems, (ii) processing of user/agent experience in information systems that can serve either purely informative purposes or form basis for higher level reasoning, e.g. concerning probable agent behaviour or estimated future experiential states. In the former case the branch of KR can provide methods and techniques of knowledge representation which can be adequate for experience representation. In the latter case for an information system to be able to give intelligible answers on questions related to user experience it has to have proper knowledge collection, storage and processing mechanisms in place that will be applied to reason about user/agent experience.

The information system of above-described qualities that in addition provides the function of decision-support is a particular type of information system referred to as expert system (ES). Expert system is a computer programme that applies knowledge and reasoning procedures as to solve problems that require human experience (of an expert) gained during a longer period of professional activity in a given domain (Rutkowski, 2005, p. 7).

The construction of expert systems is the domain of knowledge engineering, which focuses on topics such as: knowledge discovery, knowledge structuring and processing, designing and selection of inference methods, and design of user interfaces (Rutkowski, 2005, p. 8). One of the implementation areas for knowledge engineering are customer experience management systems, which role is to support organisation in the optimisation of the experience gained by their customers in relation to products, services, brands and other consumption-related entities. Customer experience management system provided the context for our early work on experience representation reported in two papers on which more will be said in chapter 4.

Returning to the main thread concerning it must be noted that advancements in development of information systems, expert systems in particular, would not have been possible without advancements in computing technologies and AI methods and techniques, as the complexity of real-life decision problem solving requires not only sophisticated algorithms, efficient data collection, storage and processing, but also enough computing capacity allowing for handling massive amounts of data, which is indispensable for modelling highly complex phenomena such as cognitive processes, knowledge and experience. The remaining part of this section will be dedicated to current developments in computer and information science that are important from the viewpoint of thesis objectives.

The claim that we live in the age in which information science is celebrating its great triumph seems easy to justify, so that it reads almost trivial: (i) we live in times where information growths exponentially in volume; (ii) information storing, processing, and access technologies are advancing form a day to another and ICT

sector has become one of the dominant industries of today, (iii) computers and computing capacity is cheaper and more affordable than ever; (iv) computing remains one of the most popular fields of tertiary education ("Education at a Glance 2011: OECD Indicators"). This list could continue for long. At the same time some good arguments can be made for envisaging disruptions in the development of both information science and ICTs, of which the most important are: (i) we are approaching the technological and physical limits of silicone in increasing the computational capacities of chips two which parallelization constitutes only temporal solution (the end of Moore's law?); (ii) the failure of classical computing based on binary algebra – the very theoretic underpinnings of information science – to deliver the promise of artificial intelligence; (iii) failure of classical computer and information science to cope with the complexities of many real life problems, in particular those related to human action and social interactions, and biological processes at large.

The foundations of modern computer and information science where laid down by theories of Turing (Turing, 1936; Turing, 1950) and Shannon (Shannon, 1938; Shannon, 2001). Turing came up with a definition of what is currently known as a *Turing machine*, a machine that performs calculations using only two types of symbols, and later co-authored the Church-Turing thesis stating that any problem that has an algorithmic solution can be solved on a Turing machine, which led to the invention of digital computers as we know today, with architecture proposed by von Neumann (Von Neumann and Morgenstern, 1944), composed of hardware (Turing machines) and software (programmes run by Turing machines). Shannon in turn came up with an information encoding method based on binary algebra that paved the way for information digitalization, i.e. representation of information in the form of strings of binary values (0,1). These inventions underpinning modern computer and information science provided for the dualistic character of computing and semantics lacking information processing machines which influenced significantly many disciplines far beyond electrical engineering for which field they had been initially meant for (compare (Guizzo, 2003)). The most surprising their influence on philosophy, to be more specific philosophy of mind. The analogy between computer and human brain had driven some philosophers to the revelation that the mind is to the brain as the software is to the hardware, a philosophical stance known as computer functionalism, or Strong Artificial Intelligence (Searle, 2004, p. 45). In other words brain provides "biological hardware" for executing programmes that are mental phenomena. This shows how profound was the impact of Shannon's and Turing's inventions on science reaching far beyond technical disciplines. Effectively, computer functionalism became a highly popular philosophical theory of mind in the second part of XX century and is still appreciated by many scientists and enthusiasts of strong artificial intelligence, becoming one of the most important physicalist, identity theories on the mind-body problem.

This topic will be looked into in more detail in chapter 3, however it is important to note now that many contemporary philosophers reject computer functionalism and materialism at large (Nagel, 1974; Jackson, 1982; Block, Flanagan, and Güzeldere, 1997; Kripke, 1980; Searle, 2004) considering it a mistaken theory. This is a very good example how, otherwise great invention of Shannon, Turing and von Neumann, led to mistaken far-fetched conclusions that has led to disappointments with regard to artificial intelligence based on classical information theory and potentially lies in the source the cause of the crisis the computer science is facing now. The below paragraph will briefly present main arguments given for why classical computer science may be considered in crisis nowadays.

A fundamental difficulty encountered by modern computer science and information technology can be found at the very fundamental level both in computer hardware and computer software. The very source of these difficulties is complexity of real-life problems which they try to solve. The complexity has two facets. One is related to the problem structuring, that involves among others: definition of the objective, conditions of satisfaction and algorithm for finding the solution, the second is the complexity of a computational problem, i.e. a task that is in principle amenable to being solved by a computer. Computational problems, which belong to the class of *well-defined* problems, i.e. problems for which a computational algorithm for finding the solution can be defined, can have different levels of complexity indicated by time needed to solve the problem, and measured in the number of states required by the algorithm to solve the problem. Problems that cannot be solved algorithmically at all, termed *ill-defined*[2], by definition fall beyond the problem-solving capabilities of computing machines (compare (Simon, 1973)). The central difficultly in problem solving is about that nearly every real-life problem is ill-defined, and even if it is possible, via relevant techniques, to structure or reduce them into well-defined computational problems their reduced complexity would still fall into the NP complexity class, which means these problems cannot be solved in a polynomial time applying known algorithms on a Turing machine.

So far the predominant approach to this challenge was to rely on the Moore's law by investing in inventing more and more powerful computer chips by increasing the number of transistors packed into a chip, and developing multi-core processing units that use parallel-computing as leverage for increasing computing speed of digital computers, the modern implementations of a Turing machine. This approach however has its limits that result from: (i) physical limits of silicone-based transistors. [3] (Lloyd, 2002), (ii) limits to parallel computation (Greenlaw, Hoover, and Ruzzo, 1995) and limits of quantum computing (Aaronson, 2007). Evidently, modern science faces

[2]Other common terms for these type of problems are: ill-structured, fuzzy or wicked problems.

[3]Meanwhile Moore's law states that transistor density on integrated circuits doubles about every two years, we are quickly (exponentially) approaching the size of atoms which is a fundamental barrier that is expected to be reached within a decade approximately

limitations of digital technology and must search for new paradigms in information processing.

Surprisingly, many of the problems that cannot be solved by computers are relatively easily coped with by living organisms, individually or collectively, regardless these are simple entities (single cells) or complex organisms like human being. Which is why bio-inspired computing is one of the possible ways out to follow (Kelso, 1995; Mange and Tomassini, 1998). Furthermore, an important characteristic of real-life problems of high complexity is that they are non-deterministic. Consequently, one of the approaches to these arising challenges is to revert to non-deterministic computing. In 2011 a large research programme was proposed under the EU's Framework Programme FET Flagship scheme, entitled "Phoenix", which goal was to elaborate a coherent, feasible and correct scientific foundation for systems that can deal with non-determinism and NP complexity through their self-organising and emergent capabilities[4]. Scientists behind Phoenix initiative claim that:

> "It [non-deterministic computing] was never of interest before as the deterministic machine seemed to be sufficient for the problems of future generation. Due to the non-deterministic nature of these problems, no current or future computer system will be able to provide sufficiently accurate results in finite time without a substantial paradigm shift in computing."

This shift is expected towards non-deterministic computations, self-organisation and emergent intelligence.

The turning point at which computer science, more precisely the Information Technologies (IT) sector, has found itself was pinpointed by Nicholas Carr from the completely different angle: application in business practice. Carr's popular paper "IT doesn't matter" published in Harward Business Review (Carr, 2003) started a more widely spread public debate on the future of IT noting that IT had transformed from a source of competitive advantage to standard cost of doing business. Envisaging what he coined the commodification of IT Carr spearheaded a spectacular campaign against mainstream belief that prosperity of IT sector is ensured. Commodification of IT according to Carr is about a shift for IT from being an asset that companies own in the form of computers, software and related components to being a service that is purchased from utility providers (Carr, 2005, p.64). It is prompted by the very similar inner socio-economic forces as those that made electricity a utility, i.e. economies of scale, standardization and ubiquity, enabling technologies (virtualisation and web services in case of IT), reliability. This view has been widely criticized by scientific

[4]Phoenix FET Flagship was a research programme proposal outlined in a CSA project filed in response to a call published by the European Commission which is classified as confidential, for which reason a reference to a printed publication cannot be provided. The author took part in the elaboration of the project proposal in cooperation with the group of researchers from research organisations from across Europe including ERCIM and Sintef.

community, as far-fetched and over-simplistic, on the grounds that computer science and application of IT will still for long time serve as source of business innovation as it supports and integrates more and more with each organisation's business processes which embody its competitive edge (Smith and Fingar, 2003), hence although certain IT services will indeed turn into commodity others never will. Nevertheless Carr's observation doubtlessly has brought to the attention of wider audience, in particular to executives deciding on IT investments, that IT has reached certain maturity phase which no longer positions it in its entirety in the topical area of innovation. It appears however that the key underlying reason for the steady disbelief in classical IT is the failure to deliver the promise of artificial intelligence which has been expected by science fiction writers, thinkers and futurist for more than 60 years already, more less the same time that we are operating under the classic paradigm of a Turing machine. Noteworthy, despite incredible advancement in engineering techniques, which allowed for building incredibly fast and reliable Turing machines (modern digital computers), there has been no fundamental breakthrough that could bring us closer to the vision of intelligent machine. Likely the classic paradigms are not apt for this challenge. Some thinkers are still optimistic however. Kurzweil proclaims the eventual occurrence of technological singularity, i.e. a point in history when technological progress becomes so rapid that it makes the future after the singularity qualitatively different and harder to predict, which according to Kurzweil will be initiated once self-improving artificial intelligence (super intelligence) comes into existence, for around the year 2045, however the development of a machine capable of passing the Turing test is expected as early as in the 2020s (Kurzweil, 2005).

Philosophers and AI scientists doubt however if artificial intelligence based on Turing machines is possible at all (Searle, 1982), not mentioning attainment of this goal in 10-20 years form today. This may result from the very nature of intelligence that is a characteristic of biological beings, and doubtlessly is underpinned by non-deterministic phenomena. Searle binds it to the free will that must have its roots in the only non-deterministic phenomena in physical world we know about: quantum mechanics (Searle, 2001; Searle, 2008), however the links between allegedly non-deterministic processes at subatomic level and the free will are yet to be tracked. Noteworthy, Einstein distrusted theories of quantum mechanics as he deeply believed that causality lies at the roots of existence, and quantum theories are merely approximations of causal processes yet to be discovered (Einstein, Podolsky, Rosen, et al., 1935). Evidently fundamental subatomic phenomena that may be at the source of intelligence are non-deterministic however the character of this underpinning non-determinism is alike that of freedom, rather than that of randomness.

To conclude, despite over 60 of development of modern information science and computer science there is a still vivid need to propose enhancements within mainstream computer and information theories within current paradigm as well as searching for new paradigms that would allow for overcoming the underpinning

limitations of the current one, especially as the information resources of today are not only reach but more easy to access than ever before, which constitutes an opportunity for many real-life applications.

Clearly digital information processing has not solved the most urging problems humanity faces, for instance prediction of natural disasters, controlling economic fluctuations, emulation social processes in security and crisis management situations. There is evidently room for further improvement and advancement in information processing technologies, with the objective to build up an information system that is capable of replying intelligently to complex queries and supporting complex problem solving.

This improvement can follow one of the two paths: (i) small incremental improvements both in computing hardware and information processing systems (software) based on old paradigms, or (ii) disruptive breakthrough caused by redefinition of classical paradigm, e.g. emergence of non-deterministic information systems. Meanwhile both directions are interesting to pursue the former appears more stable to rely on in the short run. This is also the path followed by this thesis. The incremental improvements in existing information systems can again be achieved by following two different strategies: (i) simulation of non-determinism at the level of computer software "quasi non-determinism" via established methods such as rough set theory, fuzzy logics, etc. or (ii) following the 'black box' behaviourist belief that regardless the non-deterministic character the observable behaviour can be modelled using conventional computational capacities and via modelling trying to master the complexity of problems both at the level of problem structure as well as computational algorithms.

To follow either way improved methods for knowledge representation that embrace complete mental phenomena (representation of experience) are need. The links between mental phenomena and purposive human action must be tracked and new meaning of rationality introduced to the IS design thinking. This thesis proposes to contribute to this line of research.

1.2.2 Affect and its role in understanding, defining and modelling rationality

To be able to address a fundamental question: why the improvement of information systems capable of representing human experience is a valid objective at all it is necessary first to place the problem of this thesis in a wider context. This context is provided by the phenomenon of human rationality, or in other words the study of purposive human behaviour. Adequate emulation of human behaviour remains one of the unachieved yet most desired goals of information science, artificial intelligence (AI) field to be more precise. AI both aims at creating machines capable of intelligent behaviour as well as devising systems capable of predicting, emulating intelligent

behaviour typical of biological beings. The study of human behaviour is therefore absolutely critical for this branch of information science as it builds artificial systems with required capabilities based on behavioural theories procured by social and natural sciences. In this section we will discuss the dominant theory of human behaviour that is still heavily influencing information system designers while confronting implacable criticism from contemporary cognitive sciences, cognitive neuroscience in particular. This fact has been the principal motivation for the objectives of this thesis and constitutes one of the key challenges addressed by it.

Action can be perceived as process that is not causally deterministic, but dependant on an entity acting as a decision maker, to be referred to *(rational) agent* that can be either *natural*, being a living organism, or *artificial* being an entity construed in an non-natural (non-biological) system, in our case an information system. For an action to be considered voluntary an agent must be able to exercise upon its free will. By the act of agent's free will certain qualitative endowment (input state) is transferred accordingly into an output state. The desired output state could be referred to as agent's *goal*. For comprehensive terminological discussion on *action* and *rationality* in general see Chapter 2.

Historically the theory that deals with purposive human action, which is dictated by achieving *goals* in the most efficient way is *praxeology*. The term praxeology was used for the first time to name the scientific discipline dealing with the principles of human action by Alfred Espinas in 1890 (Espinas, 1897). This embraces two basic concepts that at first will be looked at separately: purposive human action and efficient attainment of goals.

In modern times the purposive human action has been primarily the domain of interest of economics, which is probably why the best early statement and analysis of the purposeful human action can be found in writings of an Austrian economist Ludwig von Mises (Mises, 1949), which largely shaped the current economic thinking in this respect throughout XX century:

> "Human action is purposeful behaviour. Or we may say: action is will put into operation and transformed into an agency, is aiming at ends and goals, is the ego's meaningful response to stimuli and to the conditions of its environment, is a person's conscious adjustment to the state of the universe that determines his life."

Taking for the moment the claim that humans act purposefully as an axiom, we conclude that all human activity has a purpose, which means it is dictated by the content of the goal set. Goal as suggested earlier could be defined as state of the universe as desired by an agent. Von Mises claimed to this regard:

> "The ultimate goal of human action is always the satisfaction of the acting man's desire. There is no standard of greater or lesser satisfaction other

> than individual judgements of value, different for various people and for the same people at various times."

Desires has been widely accepted as the main drivers of human activity, however some philosophers of mind rightly stress that desires are not the only intentional states on which agents act directly upon, as commitments and obligations are equally important effective motivators (Searle, 2002). Nevertheless we can assume at this stage that human action is driven by reasons constituted by goal oriented motivators. We will revisit this preliminary assumption later on to show that *affect* is the underlying and primordial motivation factor.

Influence of economic thought on the study of the nature of human behaviour has had enormous impact on rationality models adapted by information science, and still has. Contemporaneously these models has been aligned with instrumentalism, i.e. the view that all reasons for action are means-end reasons, which is according to Vogler a dominant believe in contemporary philosophy (Millgram, 2001). Under instrumentalist approach an agent has a goal set dictated by a set of motivations (desires), on top of it she has a set of believes that determines the set of available means and finally a reasoning (deliberation) apparatus which allows her to match goals with adequate means relying on believes. A textbook example of a instrumentalist rationality model is the BDI (Believes, Desires, Intentions) model of rational agency, reviewed in Chapter 5.

Importantly human action is driven not by a single goal but many, often working in parallel with trade-offs between each other, however with limited means available. This is where deliberation and choice comes in. Consequently agents have to choose alternative courses of actions that will allow them to attain selected goals on minimal expense of all others, which brings in yet another concept: efficiency. Efficiency intuitively defined is the capability of an agent to take action in a way that allows them to attain goals in the most optimized way, i.e. in shortest possible time, with minimum involvement of resources and at the minimum expense of all other goals that are in a trade-off relation with those attained. Efficiency of human action, intrinsically bound with theory of choice, is the central focus of praxeology. Von Mises was a praxeologiest himself stating that:

> "No treatment of economic problems proper can avoid starting from acts of choice; economics becomes a part, although the hitherto best elaborated part, of a more universal science, praxeology."

It is particular about praxeologists that they pay a lot of attention to practical aspects, which links to the prescriptive character of propositions of praxeology. The very objective of praxeology is to identify principles of conscious, purposive human action that is driven by efficient pursuit of goals, and consequently to formulate theorems and to give advice that have practical value, i.e. that facilitate identification

and elimination of sources of the inefficiencies in purposive action. Typically, as all other disciplines the praxeology deals also with taxonomic and terminological issues related to action and investigates reasons for inefficiency occurrences. [5]

Two critical consequences arise from the approach to purposeful human behaviour as adopted by classical economists, of which the above-quoted framing by von Mises is a good exemplification, and later underpinned by instrumentalism: (i) separation of goal and motivation from deliberation process, (ii) subordination of deliberation to efficiency, search of global optima in particular. The first consequence led to rational vs. emotional dualism which at the end marginalized feelings and emotions in favour of abstract constructs such as utility, preference and goals. The second consequence resulted in placing the point of gravity on formal methods ensuring optima achievement becoming more and more abstract and distinct from the deliberation process in which a biological being involves.

Thus, originating form classical economics, largely influenced by utilitarian philosophy and praxeology, in particular within the branch of analytical economics, the classic model of rational decision-maker has emerged and traditionally has been applied for analysis of human behaviour and underpinned the classical economic theory of choice. This model is variously called, John Searle calls it simply *The Classical Model* of rationality (Searle, 2001), Edwards, von Winterfeldt and Miles (Edwards, Miles, and Von Winterfeldt, 2007) *The Rational Decision Maker*, but classically in the economic literature the term *homo oeconomicus* or *The Economic Man* are commonly used (Ingram, 1888; Edwards, 1954).

The complete presentation and criticism of this model will be provided in Chapter 5, at this point let the basic assumptions of the model be looked at: 1) the decision maker while making choices always maximizes his utility (prefers more than less), in which he is consistently logical, in the classical meaning of the term, which results in, among others, transitivity of preferences (mathematically represented as binary relations); 2) in case of uncertainty, in other words once confronted with risky choices, economic man applies probability theory to evaluate the utility of available options in which again he is logically and mathematically consistent so that the sum of estimated probabilities of the available options equals 1.

This model operates in the economics for nearly 200 years since times of John Steward Mill, and has always raised controversy. Despite having been criticised since its very early days (Ingram, 1888), it is still strongly present in the economic discourse. This can be well illustrated by quoting and excerpt for a recent publication (2007) "Advances in Decision Analysis" co-authored by the most important theorists of the decision of the second half of the twentieth century: Edwards, Miles, von Winterfeldt, Keeney, Raiffa and others:

[5]Noteworthy, the most prominent Polish praxeologiest was Tadeusz Kotarbiński (Kotarbiński, 1955; Kotarbiński, 1957; Kotarbiński, 1965; Kotarbiński and Szaniawski, 1972), and acknowledge interpreter and maven of his work, his disciple Tadeusz Pszczółkowski (Pszczołowski, 1967).

> "For the purpose of this book, *Rationality* will be interpreted as *Bayesian Rationality*, with an emphasis on (1) decision making guided by maximizing subjective expected utility, and (2) the importance of information and the processing of that information through Bayes' theorem. We take a pragmatic view of why this position is compelling. By pragmatic, we mean that systematic and repeated violation of these principles will result in inferior long-term consequences of actions and a diminished quality of life."

The above statement precedes a complete list of classical assumptions of the rational decision-maker model.

Initial criticism of the classical model of economic man's rationality, as in the case of Ingram, was held on the grounds of ethics. The concept of homo oeconomicus was attacked as a false archetype of a human being that implies people are extremely selfish and rationally egoistic, i.e. acting primarily in their self-interest. Much later, in the mid XX century, the ground-breaking work by Herbert Simon and other contemporary choice theorists spearheaded what may be referred to as the behavioural revolution in classical theory of choice. This work prompted a widespread recognition that the psychological predisposition of a human being are not well taken into account by the model of economic man, as humans simply do not decide the way the classical model imposes. Simon identified the two main classes of decision problems: well-defined (well-structured) problems and ill-defined (ill-structured) problems, of which the former, unlike the latter, comply with the classical modelling assumptions. In consequence, from then on the efforts of decision analysts confronted with ill-defined problems have been concentrated on developing appropriate methods for structuring the ill-defined decision problems, so as to translate them into well-defined problems and be able to apply mathematical analysis for finding the problem solution.

This approach and the classical model of rationality at large has meet radical criticism quite recently form the philosophy of mind (Searle, 2001) and neuroscience (Damasio, 1994), which is of fundamental importance to this thesis.

The American philosopher John R. Searle stated explicitly that the classical model of rational decision-maker is incorrect and leads to many seminal consequences, misunderstandings and wrong decisions at the micro and macroeconomic decisions (Searle, 2001, p. 9). What is important is that Searle is not criticizing rational reasoning as such, on the contrary, he shows that the classic model operates with an incorrect understanding of rationality and proposes an alternative model of voluntary rational decision making based on theory of mind intentionality, claiming that it reflects much better the actual way in which humans make choice while taking action. The main axis of Searle's criticism goes along the following points. Firstly, the classical assumption saying that human action is driven only by *desires*. Searle notes that there were other important motivators for action such as commitments and obligations,

dismissing at the same time an argument that rational agents act indirectly upon the desire to meet one's commitments or obligations respectively. Secondly, he redefines the model decision making process, rejecting utility maximization as the sole and ultimate objective of a decision maker. Instead he proposes a model of *Total Reason* being a set of statements constituting valid reasons for an agent to take given action. In addition, Searle examines in detail the relationship between decision and action, including the *intention-in-action* and *reasons-for-actions*, drawing important conclusions about the role of the conscious self in ultimate deciding about taking an action, which will be discussed in detail in chapter 3 and 5.

Searle's work is very important for several reasons. First, he showed that it is possible to talk about human action and decision making in terms of formal logic without resorting to simplifying models and assumptions of classical analytical decision analysis. Secondly, the proposed an analytical framework for analysing decision problems based on the theory of intentionality, which is an alternative approach to modelling decision making that could be adopted for representing these type of problems in information systems. Secondly, he rightly framed the problem by trying to systematically describe how human agents do in fact take decisions unlike earlier mainstream approaches in decisions analysis focused on how human agents *ought to decide* to be rational, in accordance with the classical meaning of the term rationality under utility theory. It therefore constitutes an important contribution to better methods of modelling of micro-and macro-economic processes that are so important from the point of view of the design, implementation and evaluation of socio-economic policies and the understanding of the processes that shape the social space, economic and cultural. Lastly, he casts different light on the understanding of rationality, rightly noting that it is not that theories that reject the classical model are irrational, but the understanding of rationality by the classical model proponents is mistaken. However, Searle remains constrained by intentionality dependant entirely on language and therefore does not embrace non-linguistic forms of intentionality and therefore fails to address aptly the role of feelings, i.e. subjective qualitative states of mind in decision making, meanwhile elsewhere (Searle, 2008) admits that qualitative character of consciousness and study of non-linguistic forms of intentionality should define the philosophy of mind research agenda for the ears to come.

Another revelation that is of fundamental importance to our argumentation is the one derived from the findings of neuroscience of human and animal behaviour and decision-making. As far as human decision making is concerned the ground-breaking has proven the work by the neuroscientists Hanna and Antonio Damasio (Damasio, 1994) which helped to better understand how emotions and feelings shape the decisions of a human being, providing very convincing evidence stemming from neurological brain studies performed with aid of magnetic resonance brain imaging (fMRI) technique. The main conclusion from this work is that consciously accessible reason and knowledge govern decisions only to a certain extent, whereas feelings,

emotions as well as thought processes that take place in the unconscious, or even biological states in the rest of the body make up for the rest. The conclusions from these research argument against the dualism: reason vs. emotion and feelings, clearly showing that both these spaces interact, and what is more important they show that dysfunction of brain centres responsible for one or the other leads to disturbances in the entire apparatus responsible for decision making in humans. Interestingly, Damasio concludes that there is no rationality without both lower brain functions as well as even lower bodily functions as: "the mind exists in and for the whole organism" (Damasio, 1994). Decisions depend on the emotions and feelings, surprisingly, emotions appear first, then comes the rational assessment of the decision situation in the frontal cortex of the brain. Furthermore, in a situation of strong emotions actions are taken automatically without the involvement of brain centres responsible for conscious reasoning. Most importantly of all, the lack of emotion can lead to serious dysfunctions in decision making, i.e. inability to make good decisions nor taking them at all, which is confirmed by clinical cases of patients with brain damages or other brain dysfunctions in areas of the brain responsible for emotions. These patients, despite normal capabilities of rational overview of the decision situation, for instance identification of relevant alternative courses of action or consideration of the arguments for and against each of them, are not able to eventually to chose and pursue any of them.

Another important point made by Damasio is that it is untrue to say that emotions and feelings are ephemeral, since they can be studied and explained just in the same way as other functions of the brain and body [ibid., p. 10). This gives hope that the emotions and feelings can be described using the language, natural or artificial which would allow their representation in information systems. Emotions and feelings can be illustrated by a biological brain states and their counterparts around the body. It also shows how complicated a mechanism we confront here, the decisions not only depend on conscious rational assessment but also emotions, feelings and states of the whole organism governing the decision-making processes in the unconscious.

Later (Damasio, 1999), and in a very recent (Damasio, 2010) work Damasio substantially contributed to the theory of emotion and human behaviour by clearly telling apart emotions and primordial bodily feelings and distinguishing their different role in shaping human behaviour. His work together with that by affective neuroscientist Jaak Panksepp (Panksepp, 1998) who was one of the originator of the idea of primordial feelings and Joseph Ledoux who disentangled the neuronal precesses that stand behind emotional learning form a modern neuroscience foundations for building theories of human conscious experience, behaviour and decision making. This work will be reviewed in Chapter 3 as a basis for developing information affective model of experience in Chapter 4.

This is how we arrived to the relation of the issues discussed throughout this section with the thesis objective. The objective of this thesis is to work out theoretical framework for building methods of human experience representation in experts systems supporting efficient emulation of human purposive behaviour, i.e. making decisions by humans. We believe that systems that emulate human behaviour cannot do without proper representation of complete mental phenomena that orchestrate the purposive action process as depicted in the purposive behaviour efficiency improvement cycle. These mental phenomena include not only knowledge, which is traditionally considered the only relevant mental content in information science, but also subjective components of experience dependant on emotions and feelings. The theoretical framework allowing for representation of the complete plethora of mental phenomena influencing purposive behaviour of natural agents is required for information systems to be able to emulate and predict such behaviour, which in turn is need for applications that range from highly usable and adaptive information retrieval systems, societal processes modelling to creating believable artificial virtual characters, and machines (robots), and solving complex societal problems.

Now, let us elaborate on how experience and its nature relates to purposeful behaviour, hence decision making. It is impossible to investigate the nature of experience and its relation to agent's voluntary action without referring to consciousness. Accepting that a leaving creature with fully functional brain is always conscious, only moving between the three main conscious states: wakefulness, dreaming and dreamless sleep, and observing that there is no experience without being consciousness, i.e. being alive, and that whenever one is conscious one experiences, it becomes clear that consciousness equals to experience. Whichever observations, qualities and rules that may be applicable to consciousness can be directly applied to experience. This is basically the fundamental reason why the nature of consciousness will be reflected upon in the thesis.

Consciousness is however the most complex and difficult phenomenon to analyse of all introduced so far, for one primary reason: it is a phenomenon that is intrinsically subjective, the access to consciousness is only possible via introspection, which is a deficient and scientifically unreliable mode of cognition because in introspection no difference can be made between the object and the subject of cognition. Furthermore consciousness is at the same time material, i.e. based on the biological processes in the brain, and irreducibly mental, for which reason it is impossible to reduce consciousness to any physical process that could be fully described only in the third personal terms (Searle, 1999).

The detailed discussion on the nature of consciousness will continue in chapter 3, now let us only consider the character of consciousness as proposed by John Searle as to better illustrate how the philosophical debates on consciousness influenced the model of rational experiencing agent.

Consciousness, according to Searle (Searle, 1999, p. 120), has ten principal features. Firstly, which has been said already, consciousness is subjective. Secondly, consciousness is unitary both vertically and horizontally. Horizontally because the conscious states are at any point in time experienced holistically within one unitary field of consciousness, in other words we experience all stimuli that reach us at one time as one conscious state. Vertically, because it is also consistent in time for which reason it needs memory, at least short term memory. Furthermore, consciousness is intentional, i.e. it has this particular capability to be about or directed at elements of external world [6]. Fourthly, each conscious state is accompanied by a particular mood. Next, normally the conscious state are structured, i.e. their are fit in a bigger picture, wider context, by the mind, which is well illustrated and explained by Gestalt psychologist. Moreover, within the spectrum of consciousness there are focal areas governed by attention. Furthermore conscious states are normally located in the wider context such as time of the year, geographic location, weather condition, etc. appreciated by the mind. Next, an agent is familiar to experienced conscious states to the greater or lesser degree. Moreover, conscious states have this unique capacity that they refer to entities beyond themselves, as one thought recalls another, in this sense consciousness is always active. Finally, there is a subjective qualitative feeling associated with each conscious state.

All above characteristics of consciousness are highly relevant for understanding the way conscious agents undertake action. Meanwhile, the qualitative aspects of conscious experience have been so far largely neglected by classical choice theorists and mainstream economics. However, it has been already acknowledged that information systems simulating customer decisions fail to deliver meaningful outputs, for which one of the reasons is that these systems do not include variables representing the qualitative character of consciousness, therefore experience, which reaches beyond the scope of traditionally defined knowledge.

To conclude, as this thesis looks for the ways of representing experience in information systems that could support better understanding of human decisions it is indispensable to investigate in more detail the phenomenon of consciousness, as it provides the epistemic ground for voluntary, purposive action of any agent, a human being in particular. In particular it is relevant how the unity and complexity of conscious mental phenomena embracing: objective knowledge, subjective affect, freedom of will, memory and learning and unconscious processes could be represented in information systems.

[6] Intentionality will be discussed in detail in chapter 3

1.2.3 Methodological inspirations: Leibnitz's *Calculemus!* and Wilber's *orienting generalizations*

The previous sections has clearly showed that the problems to be addressed by this thesis are intrinsically multidisciplinary. Problems of human action has been of interest to many disciplines: philosophy, psychology, anthropology, neuroscience, economics, management sociology, information science, to name a few. However historically these problems has been looked at from very different perspectives as imposed by each discipline.

History of science shows that often conflicting ideas of scientists lead to stagnancy in the progress of science or waist of scientific effort. However, there is a noble movement in the philosophy of science to which spirit this thesis would wish to contribute. The very essence of this spirit has been aptly verbalized by Leibniz and his famous call: "Calculemus!", i.e. "Let's calculate!". Although this Leibniz's thought is most commonly associated with the beginnings of analytic philosophy, early efforts to find suitable symbols for the representation of mental phenomena, it can also be interpreted more widely, so as Leibniz would urge his fellow scientists to terminate endless disputes and make effort to speak the same language in the effort to understand the true account of things (compare (Heller, 2008)). This can be perceived as precursory to the currently widespread multidisciplinary approach in science which brings value to the discussion and facilitates scientific progress.

The good historic example of fruitful interdisciplinary thinking can be credited to Herbert A. Simon and his colleagues, who worked towards bringing together the insights on human behaviour form psychology and economics, the two disciplines that used to have been at odds before Simon's ground-breaking publications in mid XX century.

A recent important contribution to this movement must be credited to American philosopher Ken Wilber, the founder of integral school of thought who enriched the philosophy of science with a method referred to as "orienting generalizations". His method stems from the believe that it is seems impossible that a human mind could be completely wrong at all times:

> "I don't believe, that any human mind is capable of 100 percent error. So instead of asking which approach is right and which is wrong, we assume each approach is true but partial, and then try to figure out how to fit these partial truths together, how to integrate them-not how to pick one and get rid of the others." (Wilber, 2001)

Therefore Wilber's method is about assembling all the orienting conclusions from all the fields which doubtlessly have important truths to contribute. Then these truths shall be arranged into chains or networks of interlocking conclusions and finally once the overall scheme is developed that incorporates the greatest number

of orienting generalizations, the partial nature of most narrow approaches shall be subject to constructive criticism, importantly it is not the truth but its partial nature that is to be criticized (Crittenden, 1997).

Noteworthy, the great affirmation, and to some extent the materialisation, of the Leibniz's call "Calculemus!" of nowadays is Wikipedia[7], an on-line collection of over 3,5 million (counting only English version) collaboratively created articles, founders and contributors of which we owe sincere acknowledgement.

1.3 Thesis proposition

The central proposition of this thesis states that experience, broadly understood as the entirety of contents and quality of a conscious mind state, can be satisfactorily represented in information systems, and any information system which objective is to emulate natural agent behaviour with satisfactory faithfulness cannot do without a sound experience representation framework. To achieve this it is necessary to realize and accept, based on convincing evidence from neuroscience, that the missing subjective component of experience is *affect* that forms and integral part of natural agent's experience, and determines, or at least impacts profoundly the behaviour of natural agents. Relating affect to knowledge would result in a satisfactory approximation of experience. It is to realize as well that the subjective dimension of experience, classified as affect, is not entirely private, subjective epiphenomenal entity, as some proponents of qualia suggest, but rather as Panksepp proposes can be studied in objective terms in a satisfactory manner by looking into neurological correlates in the brain (Panksepp, 1998). By identifying affective correlates of intentional contents of states of mind, which build up knowledge, we can exploit a broader concept *experience* for the purpose of more accurate emulation of natural agents' thinking process and behaviour in information systems.

Consequently, agent experience should be regarded as composed of conscious states of mind (intentional states) that have both an objective component: the intentional content, and a subjective component: the affect, both playing a pivotal role in determining voluntary action of an agent.

Assuming the unity of conscious experience we argue that treating intentional contents, i.e. knowledge, and affect, i.e. subjective feelings to intentional states, separately is mistaken. We propose the way how these two seemingly different mental phenomena could be represented in a unified way in an information system.

The above overall proposition entails the following related propositions:

1. There is a causal relationship between experience and purposive action.

2. Mainstream information systems supporting emulation of natural agent's behaviour fail to create believable artificial agents due to the lack of robust

[7] `www.wikipedia.org`

frameworks for experience representation, in particular fail to aptly capture the affective dimension of experience.

3. Experience, including its subjective component, can be modelled and represented in information systems.

4. Despite certain differences (discussed by this thesis) between experience and knowledge, regardless these differences are merely consequences of differences in understanding these two terms by philosophy and computer science, it is possible to adapt and apply knowledge representation and processing methods for experience representation and processing, specifically with regard to emulation of natural agent's behaviour in information systems.

1.4 Thesis structure

In total, this thesis is composed of six chapters. This introductory chapter will be followed by five chapters which are briefly summarized below, including a final chapter with concluding remarks.

Chapter 2

In chapter 2 we will introduce key terms used in the thesis. This will allow us to set up terminological background indispensable for subsequent line of discussion. The terms introduced in this chapter will include among others: information system, knowledge and experience, knowledge and experience representation, process, efficiency, affect, purposive behaviour, decision and rational agent.

Chapter 3

The objective of chapter 3 is to set up the theoretical, mostly philosophical and psychological, basis for further discussion on experience representation in information systems. This is to provide solid epistemic background for formalisms to be introduced in further chapters regarding purposive behaviour and experience representation. First, the nature of experience will be analysed in more depth, relying on what has been initially said in chapter 2. In this chapter the human experience will be characterised from the perspective of epistemology and modern cognitive sciences. Then its intrinsic relationship with consciousness will be looked into and insights from the philosophy of mind provided. Consequently the concept of intentionality will be discussed specifically its significance for the study of conscious experience and meaning as well as how it relates to language, both natural and artificial. This chapter will also embrace a discussion on qualia, alleged subjective qualities of experience. The subject of quality of experience will be also addressed in the context of natural agent's affect and findings of contemporary cognitive neuroscience

about human and non-human emotions and other affective states. Furthermore, experience will be related to human decision making, thus human action. Although the human decision making will constitute the main focus of chapter 5, in which we will consider application of the proposed experience representation framework to modelling rational agency, already in the concluding part of chapter 3 relevance of experience to human decision making will be sought. Finally, we will report the results of a study that we had conducted in order to empirically confirm the influence of affect on rational judgement, the effect which has gained significant attention from cognitive psychologists and neuroeconomists spotting new light on the very basis of rational agency.

Chapter 4

The purpose of chapter 4 is to introduce a general purpose experience representation framework that could be of use in information systems supporting natural agents behaviour emulation and analysis, in particular systems that support emulation, analysis and simulation of human behaviour. The aim of this chapter is therefore to look into how exactly experience could be modelled and represented in an artificial, formal system.

Towards this end we will first introduce the conceptual framework for unified representation of affect and knowledge in informations systems, which will allow us to introduce the formal definition of experience and propose experience representation framework. Meanwhile the said framework will be built on the mainstream knowledge representation approaches without entering into a discussion on their validity nor efficiency, the complementing part of the model corresponding to affect will be studied in more detail. The chapter will include a review of emotional and affective models used in information science and AI with the objective to identify their main weaknesses and limitations Finally, we will be able to conclude with proposing a general purpose theory of experience applicable to representation, emulation and processing of experiential phenomena in information systems. This will feature the presentation of a preliminary version of the framework proposed in earlier publications, in collaboration with Dr Dominik Ryżko from Warsaw University of Technology, proposed in the context of consumer behaviour, to be more precise created with the purpose of modelling customer experience in Customer Experience Management systems, a type of information system used by companies to support customer relationship and loyalty management which has been developed, as well as a most recent updated version of the framework constituting a generalized reformulation and extension of the original approach.

Chapter 5

In chapter 5 we will consider how the proposed experience representation framework could enrich and enhance the existing approaches to modelling rational agency in information systems.

Experience will be put in the context of rational choice and we will undertake to analyse how our account of experience could improve the way how rationality modelling has been done and still is being done contemporaneously by economics, AI researchers and information scientists. For this purpose we first look at how historically the problem of choice has been addressed in different branches of social sciences, economics and psychology in particular, to set up proper context for the problem. This will be followed by an overview of contemporary theories and models of choice as to identify current state of the art in choice theory and challenges yet to be addressed, and thus provide arguments for the validity of the problem we address. Next we will present an overview of insights into individual choice theory provided in recent years by contemporary neuroscience, which will result in a set of human decision determinants that has until recently been largely neglected by economists. The concept of bounded rationality will be revisited adding up more behavioural effects discovered recently by brain scientists. Then we will look into the problem of choice processes representation in information systems, with particular emphasis of the mainstream contemporary rational agency model: the BDI framework. And finally we will consider how our account of experience could fit in the BDI framework and rational agency models at large and how this would contribute to the efficiency gains in agency modelling in information systems and better practical applications.

Chapter 6

The final chapter will summarize the main thesis results. This will cover the consideration of the limitations of models and methods proposed as well as will outline the further research programme aiming at addressing these limitations. This chapter will also include a subsection on ethical issues related to the application of experience modelling information systems.

Chapter 2

Terminology

2.1 Introduction

In this chapter we will introduce the key terminology used in the thesis. Many terms used throughout the thesis are problematic terms because despite decades, more often ages, of scientific disputes no ultimate, commonly accepted definitions have been worked out. For this reason the terminological discussion has been moved to this separate chapter so that the terminological disarrays do not disturb the flow of argument presented in the chapters to come. Importantly however most of the problems discussed by the thesis have roots in the terminological deliberations for which reason some of the key thoughts presented in this chapters will echo in those to follow. For the benefit of clarity some discussion may be therefore brought back and repeated later on.

The terms to be introduced in this chapter cover a wide range of scientific disciplines which is the consequence of the interdisciplinary character of the problems tackled by the thesis as well as the interdisciplinary method adopted by the author in challenging them, i.e. looking at the problems from different angles by taking advantage of different insights to the same or related problems from the perspective of different scientific disciplines.

The introduced terms can be structured in several groups and will be discussed in the following section:

- Terms related to rational agency (incl. decision, rational agent, purposeful behaviour, action, reaction, habit, scarcity of resources and efficiency, motivation, self)

- Terms related to epistemology (incl. process, knowledge, experience),

- Terms related to brain science and psychology (incl. emotion, feelings, affect, neural circuit, mental state, consciousness)

- Terms related to information and computer science (incl. information system, intelligent information system, artificial intelligence, knowledge and affect representation, information retrieval system, affective computing)

2.2 Terms related to rational agency

As we are investigating in this thesis methods for efficient representation of experience in information systems which aim is to emulate live organisms, human in particular, behaviour we will operate across this thesis with basic terms that link to the fundamental life processes: the *behaviour*. Human and animal behaviour is tightly knit with fundamental phenomena such as scarcity, choice and action which will be considered in this sub chapter.

Scarcity Scarcity and consequent necessity of choice is the central problem of economics (Schiller, 1991; Begg, Fischer, Dornbusch, 1993; Kamerschen, McKenzie, Nardinelli, 1991; Hyman, 1989; Bowden, 2002). Most of definitions of economics refer directly to the problem of scarcity. Hyman defines economics as the science of applying the limited production resources within a society in a way that can best satisfy the unlimited needs of its members (Hyman, 1989). Lionel Robins (1932) points at economics as the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses, and Kamerschen, McKenzie, Nardinelli (kamerschen1991) explicitly define economics as the science that investigates how people cope with the problem of scarcity. All the above-quoted definitions express similar thoughts, it can easily noticed that the common denominator for them is the concept of *scarcity* and *choice* that implies *decision*. Scarcity is then nothing else but an imbalance between the manifested needs and available resources that can satisfy them (Hyman, 1989).

Reismann provided an interesting account of scarcity problem, noting that needs are products of human mind, catalysed by imagination, meanwhile the goods and services that can satisfy these needs are always some real, physical entities creation of which requires a given amount of resources. Consequently scarcity emerges as a result of the confrontation of unlimited human imagination and physical limits of production (Reisman, 1998). The phenomenon of scarcity is therefore eternal and independent of human endeavours aiming at its mitigation.

It may be considered whether scarcity indeed is a problem that humanity would never be able to solve, considering it in light of the diagnosis provided by Reismann. According to classical economics the means of production include: land, work, capital and knowledge (technology) (Begg, Fischer, and Dornbusch, 1993). Contemporary economics indicates the growing importance of the latter, and diminishing labour and capital intensity of production and service delivery (Drucker, 1969; Drucker, 1993; Machlup, 1962). Taking this into account together with an observation of progressive

digitization of goods, services but also increasingly production means, taken this account an extreme by some futurists such as Kurzweil (Kurzweil, 2005) we feel encouraged to propose a hypotheses about the total or far-reaching digitization of all goods, services and processes and the transfer activity of the human mind from the physical world to cyberspace. It is worth noting that knowledge as a product of the human mind, following the reasoning by Reismann, is not subject to the limitations of scarcity, as creativity seems to have no bounds. Under the paradigm of knowledge-based economy both on the substrates and on products side of production equation there is the unlimited imagination of a human mind. In such a hypothetical situation scarcity is eliminated or substantially reduced. It is worth noting that in cyberspace scarcity is created artificially, Second Life® can be taken as example, where limits are imposed on the amount of available virtual land and other movable and immovable resources, which are otherwise nothing but lines of computer code that can be multiplied indefinitely. This trivial example shows that a world that we are all used to, a world characterised by scarcity is the only one in which humanity knows how to operate. We can risk a statement that, due to the fact that scarcity determines the market value of products and services correlated future successful business models in advanced knowledge-based economy will rely on the scarce physical production substrates that are indispensable for delivering knowledge-intensive goods or services, and the owners of this resources will dictate the condition. This is what we already observe in ICT sector where hardware or access-bound software services underpins most successful business models.

However, considering a situation in which scarcity is limited by increased role of knowledge on the production means side but not completely eliminated, which would be the case of Kurzweilian singularity, there would remain in force another very important restriction, now on the demand side: consumption capacity limits. The capacity of people and organisations to absorb goods and services, either material or symbolic, is limited which is a simple consequence of the spatio-temporal constraints.

It would therefore be reasonable to strengthen Reismann's theorem as follows: scarcity arises from the confrontation between the limitlessness of human creativity and the physical limitations of production capacity on the one hand and consumption capacity of consumers on the other. In such a case the only available way of increasing production and consumption is on the one hand an increase in efficiency of production processes and intensification of consumption on the other. This conclusion brings us to the next two fundamental problems: *choice* and *efficiency*.

Behaviour The natural consequence of scarcity is the necessity to make choices, or as Barron and Lynch put it, the problem of choice arises only if there is scarcity. Hall and Lieberman add that the necessity of choice in a consequence of scarcity, and constitutes the source of all problems of interest to the economics (Hall and Lieberman, 2001). We may extend this claim beyond economics and note that choice

efficiency is central for all behavioural sciences and evolutionary biology. The key question economics and management science in particular, confront is how organisms cope with scarcity by improving efficiency of behaviour through taking correct decisions, meanwhile evolutionary theorists, approach decision making and behaviour from the perspective of life management and try to figure out how these strategies are encoded into genes and passed across generations. Later in chapter 3 we will bring arguments for that emotions play an important role in these processes.

Choice can be defined broadly or narrowly. In the narrow meaning of the term choice is an indication of one of the options. In a broader sense, much more common in economics and psychology, the choice is a conscious mental process that results in the selection of one of the available options by comparing their value to the decision-maker. Importantly, choice always results in action, and always is about some sort of behaviour. Choice in a wider sense is therefore a complex process comprising: identification of available action options, assessment of the value of each individual option (evaluation), comparison of available options, indication of one of them and finally putting this into action. In a narrow sense, the choice is therefore only the last but one stage of the complete choice process, that is an indication of one of the options considered. The part of the process that precedes indication of the chosen option will be referred to as *deliberation*. Choice and its determinants is studied under wide variety of fields: economics, management science, psychology, social psychology (sociology), cognitive science, and biology.

As living organisms constantly confront with scarcity while carrying out their life management they constantly choose. Behaving is nothing but a life process, continuous across lifespan, which is about choosing one of the potentially limitless ways of conduct, one instance after another. For this reason across this thesis terms *action*, *behaviour*, *conduct*, *choice* and *decision* will be used often interchangeably, as we recognize that all of these are in principle different terms that represent the same phenomena at the low level used with varying preference across scientific disciplines.

Notwithstanding, behaviour is broad term that covers all sorts of agent's actions, and it is important to differentiate basic types of actions to be able to narrow down the focus of the discussion. We are short of place here to consider action typologies within a proper historical contest. Instead we claim that most of the classical theories of choice, typically elaborated by economists, has been made obsolete by recent findings of neuroscience. As a general rule the so called armchair theories of human actions failed on one important element: recognising the due role of emotion and unconscious brain processes in human behaviour. We will come back to this point on many occasions as well as we will introduce and review the classical model of rationality in Chapter 5. For the moment let us introduce a clear, and relevant to our further discussion, taxonomy of behaviour provided by LeDoux et al. (Gazzaniga, 2009, p. 905), which will be taken as reference point for the remaining part of the thesis.

Ledoux et al. propose that most of behaviours performed by people and other animals can be assigned to one of the four categories: *reflex*, *reaction*, *action* and *habit*. A *reflex* is a "stimulus-evoked response that usually involves a single muscle or a limited group of muscles". Eye closure in response to a blow of air or removal from the source of pain (for instance, a burned finger) are common examples of reflex behaviour. A *reaction* is much alike a reflex, as it is elicited by a specific stimulus, but it differs in that a reaction is a response to a stimuli by the entire body rather than a limited group of muscles. Body response to fear such as freezing that involves virtually all muscles serves as a good example. Reactions are also called *fixed action patterns* (Tinbergen, 1951), which highlights the two key characteristics of this behaviour: neurological fixation and connection with *action* that typically involves entire body. As for neurological fixation both reflex and reaction are neurologically hard-wired, which means that they are a sort of neurologically programmed behaviours that escape conscious control. They are effectuated automatically and elicited, and in this sense defined by, a specific stimuli. Reflexes and reactions are programmed either genetically, for instance rodents born in laboratory respond with fear to cat's odours (Blanchard and Blanchard, 1972), or through associative learning. The later feature makes both reflexes and reactions plastic in the sense that they can be modified by means of associative learning. As associative learning is a process that takes time, at a given moment reflexes and reactions are fixed responses. As for *actions* Ledoux characterises them as instrumental responses that "are emitted in the presence of certain stimuli that direct behaviour toward goals". As mentioned earlier actions are similar to reactions in that they involve entire body rather than specific muscles. However, the distinction between reaction and action is very clear and important as action is not a programmed response but rather a flexible response modulated by assessment of relationship between consequence of action and the goal. This brings us back to the considerations at the end of the previous chapter about purposive human action. Actions therefore are human and animal behavioural responses that can be consciously modulated in a way that they best serve the organism's goals, for which reason they are often referred to as *goal-oriented* or *purposive* behaviour. Furthermore and importantly actions must involve motivation, i.e. the reasons why an organism undertakes and, importantly, *pursues* an action. At this stage we should add a type of action not considered by Ledoux, notably a free action. From the philosophy of mind perspective a free action is that which does not have a causally sufficient conditions that determine the action. There is, as Searle proposes (Searle, 2001) , *the gap of free will* between the antecedents of a free action and *prior-intention*, i.e. the moment when the decision is taken and finally *intention-in-action*, which is actually doing what has been decided, that all in one lead to action accomplishment. The problem of free will and its existence is to be discussed in more detail in the following chapters, but here let us propose that, given there exists a genuine free will, we will refer to actions dependant on free will as *free* or *voluntary actions* and

on the contrary *involuntary action* is that which is carried out without involvement of free will. Whenever we use the term *action* with no attribute let it be assumed that it signifies any action, regardless voluntary or not. A careful reader should ask, what is then the difference between involuntary action and a reaction. Action in general involves deliberation, an assessment of available options of conduct with regard to a goal. This assessment can be done both consciously and unconsciously at the neuronal level. As free will would not exist otherwise than in consciousness the answer to the question would be that involuntary actions are those actions that are performed partly of fully in the unconsciousness. Another type of an involuntary action is that which is carried out against the free will, i.e. a *forced* or *imposed* action. Importantly however a forced action is a one in which consciousness and free will are involved, however in this special case a subject self despite agreeing to the imposed way of conduct acts upon his free will. Coming back to Ledoux's typology, the last category he considers is a *habit*, which is a type of action that in course of repetitive coexistence of a given learned instrumental response to a given stimulus becomes inflexible an automated response which has lost its connection to the attainment of a goal. A habit is not hard-wired as reflexes and reactions are, as it originates form goal-directed action, in which by routine the deliberation process is lost. Again *habits* can be changed via learning but usually it is not easy, and can lead to pathology if they lead to maladaptive behaviour. Noteworthy, a *skill* is related to *habit*, as it involves a set of trained responses to stimuli occurring when a given activity is performed.

Motivation While discussing action varieties a few paragraphs above we said that actions are *goal-related*. What is a *goal*, how goals are determined and why an organisms wants to pursue them are all questions tackled by theories of *motivation*. The study of motivation is nothing but trying to answer the key question posed in the preface: what makes living entities feel like doing anything. This question can be considered at two levels: (i) the fundamental metaphysical level: *why* living organisms what to achieve or avoid certain states of universe, why they have preference of one states over another, as well as epistemic level: how this is implemented in a living organism. As for the former we will take that it is the nature of things that living organisms are motivated or reluctant to act throughout their life and will focus more on how this is engineered by evolution, so how it is that organisms have preferences and urge or reluctance to behave. Modern brain science links motivation to the brain capacity to evaluate and select alternative ways of conduct. Some theories even speculate about specialized areas in the brain that are responsible for mediation of behaviour which corresponds to governing motivation. Influenced by brain scientists such as Damasio, Panksepp and Ledoux as well as philosophers of mind such as Searle we believe the motivation is entirely encapsulated in experiential subjectivity. The capacity of brain to assign any value to anything results directly form its capacity to

have subjective, qualitative states of mind. Consequently the cognitive and affective components of conscious experience come together in behaviour where the latter provides values needed for an organism to set goals and know when these are attained and the former allow an organism to relate states of the world to goals and figure out what is the most efficient way to attain them. Having said that we risk stating that feelings, including feelings of emotions, are the currency of behaviour that provide meaning to it and thus provide for behavioural motivation. Without subjective feelings organisms could not involve in voluntary behaviour. Decision making cannot be understood without understanding affect.

Self In order for subjectivity to exist there must be an entity that engages in subjective experiencing and behaviour, voluntary free action in particular, this entity is referred to as *self*. It is a philosophical discussion falling beyond the scope of this thesis whether self is necessary for subjective experience to emerge or rather a self is a compound of subjective experiences. The latter stance most typical of Humean empiricism is cultivated to this date by many contemporary philosophers, including Dennett who believes that self is an illusion created by stream of consciousness, conscious perceptions ordered by brain in a way that an illusion of continuity and unity is created. The former account of self is closer to naturalistic position in which biologically fostered mindful being is the primordial subject of experience, which evolved consciousness capable of decision making that resulted an evolutionary efficient apparatus of environmental adaptation, which stance is contemporaneously supported by philosophers such as Searle (Searle, 2004), but primarily accepted by biologists who like Ledoux believe that "the existence of a self is a fundamental concomitant of being an animal" (LeDoux, 2002, p. 38). Such a biological account of self stems from immunology for which defining the constituents of identity, distinguishing one organism from others, among others for the purpose of devising mechanisms of defence is the main preoccupation. Interestingly however contemporary immunology theorists increasingly accept ill-structuredness of the concept of immune selfhood and arguments coming from transplantation and autoimmunity studies questioning strict dichotomy of self and non-self. Regardless philosophical debates, which however have certain important consequences on behaviour related concepts including existence of free will, it is enough at this stage for our purposes to note that self and subjective experience are a two sides of the same coin. Ancient philosophers considered *soul* rather than *self* which are in principle two different terms corresponding to the same phenomenon. Interestingly Aristotle considered that soul is constituted by the core essence of a being, the first activity of the body, which in case of humans is the rational behaviour, thus the rational activity constituted according to Aristotle the essence of a self (Aristotle and Sachs, 2001). To put it in other words self is behaving by an organism, behaviour is determined by subjective experiencing constituting motivation and subjective experiencing constitutes self, at which point we arrive at

recurrent loop. The aim of this thesis is not however to disentangle this recurrence dilemma but rather map and logically interrelate the stages in the loop.

Regardless self objective existence or illusory nature we can accept the concept of *self* as that which identifies a particular being, building on a wide definition of self by William James who believed that a self is the sum total of all that one can call his, the sum total of who one is. After Aristotle we will treat self primarily in link to action, for which reason we will use interchangeably other terms to refer to a self as the subject of action or experience: *(rational) decision maker*, *(rational) agent*, *organism*, *animal* or *being*. Throughout this thesis we will most commonly use the term *agent*, making an important distinction between *artificial* and *natural agent*. The latter would correspond to a self of a living creature, human being in particular, meanwhile the former to a *quasi-self* of an abstract entity that emulates behaviour of a *natural agent*, a computer programme for example.

Rationality We noted earlier that the type of behaviour we will primarily concentrate on hereof are actions, i.e. responses emitted in the presence of certain stimuli that are aimed at achieving goals. We have also said that goals are motivated by subjective experiential states, therefore by feelings, which will be discussed in the following section. A behaving agent performing an *action* in order to achieve her goal involves in a mental process which is about selecting most efficient means, i.e. available courses of action leading to the goal attainment. This mental prices can be called differently: *reasoning*, *deliberation*, *reflection* or *thinking*. Aristotle commonly quoted sentence from his On Rhetoric states that "thought by itself moves nothing(...). What moves us is thought aiming at some goal and concerned with action" (Gross and Walzer, 2008) proposing that thought can be independent of action, but we believe it is not the case. Adopting the cognitivist approach under which acting is influencing environment via some kind of movement [to be checked + literature from philosophy] and given that thought capable minds evolved only in organisms able to move, presumably only those capable of voluntary movement, we assume thought even abstract is always related to some kind of action, e.g. writing, speaking, albeit may be differed in time. The mental process of which we talk about above accounts for *rationality* in broader sense. *Rationality* or *practical reason* is the general capacity of mindful creatures for resolving, through reflection, the question of what one is to do (Wallace, 2009). Rationality in contrast to practical reason may also mean not the capacity itself but acting upon this capacity. Furthermore, narrowing the sense of the term event further, it may imply acting in a particular way, i.e. acting rationally. This is when the capacity itself ceases to be the central focus of consideration, and when normative approach takes over mandating consideration of what one shall or shall not do. In such a case rational behaviour means a behaviour performed in accordance with a set of clearly defined rules.

Practical reason has always been in the centre of philosophical debate. From Aristotelian highly moral practical reasoner outlined in his On Rhetoric:

> "No one wishes for anything except when he thinks it good.",

to contemporary instrumentalist theories of rationality, largely influenced by Humean famous assertion:

> "Reason is, and ought only to be the slave of the passions." (Hume, 2003),

which accounts for that all reasons for action are means-end reasons, i.e. practical reason is an interplay of beliefs and desires within the mind, which echoes in different variations across all contemporary accounts of Anscombe (Anscombe, 1957), Korsgaard (Korsgaard and O'Neill, 1996), Bratman (Bratman, 1987), Dennett (Dennett, 1989), Searle (Searle, 2001), Davidson (Davidson, 2004).

Apart from philosophy of practical reason the biggest footprint on the concept of rationality has been left by economic theories. Economists introduced a narrow understanding of rationality by conceptualising, typically of economic method, a model of *homo oeconomicus*, i.e. the economic man, which contemporaneously evolved into so called Bayesian rationality. We will review the account of rationality by economic theory, choice theory in particular, in detail in following chapters, here let us just briefly outline the concept of Bayesian rationality after Edwards (Edwards, Miles, and Von Winterfeldt, 2007) which involves: (i) decision making guided by maximizing subjective expected utility, and (ii) the availability of information processed through Bayes' theorem. The model of rational economic man assumes first of all that the decision maker can always weakly order all states of universe that can be available to him by which he defines his preferences, as well as he always chooses the best of them as he always maximises the value that he can expect from the choices available to him. On top of it his preferences must be transitive.

In mainstream economics literature but also instrumentalist philosophy of rationality the role of subjective states of mind is limited to motivational function. In plain words affect sets the goal and 'reason' does the rest i.e. matches the beliefs with goals and urge of efficiency to sellects the actions accordingly to efficiency or evolutionary fitness maximisation principle. We will argue that it is partly true. It is true that subjective experiential states determine goals but also they directly influence the process of deliberation, and in case of reactions and habits may constitute a stimulus directly eliciting response. It is difficult to judge whether in such a case the behaviour is irrational especially as these responses may either be unconsciously derived by a given neuronal circuit, be hard-wired by evolution or trained, in which case irrationality would not be an adequate term to describe such responses. On the other hand deliberative responses may as well lead to maladaptive behaviour and may not be optimal form the global perspective.

Brain science proses terminological problems for theories of choice and practical reason as the demarcation line between that which is rational and that which is not seems to pale in confrontation with neuro especially if matched with epistemological determinism. How can a natural agent be irrational if he acts merely as a automaton transforming one set of input states into output states according to the . Therefore if one rejects free will one must also reject irrationality, but those that believe that free will is a real gap in causality of physical world fully dependant on self, how can we ever judge on whether given behaviour is irrational, how can we be sure that there where no reasons justifying the behaviour opted by the agent, especially as there might have been implicit internal reasons to which not only our consciousness but also agent's consciousness had no access.

For the above reason we suggest to reject the rationality/irrationality dualism. By nature natural agent are rational, the same way physical world is characterised by cause and effect relationships, in the sense that there are always reasons that justify agent actions although these may be sometimes implicit. However avoiding using the term rational and irrational in this thesis is simply not possible as it in large part is about the necessity of abandoning old concept of rationality. Therefore whenever the term *rational* or *rationality* are used they either mean Bayesian rationality or intuitive or common language meaning of the term, i.e. related to reason, truth logical inference rather than emotions and subjective feelings and judgements. In the latter case we will put the term in inverted commas: 'rational'. Finally we will rather use the term *deliberative* then *rational* to emphasize that there is a thought process going on in the mind of behaving agent.

For the purpose of completeness it is necessary to briefly comment on yet another term related to rationality, namely *problem*. Problem is simply the gap that exists between the current state of universe and that desired by an agent, i.e. the goal state [compare Pounds, 1969]. Obviously the problem is solved when this gap is eliminated, so as the desired state becomes reality (Bartee, 1973). In principle problem solving is yet another term equivalent to behaviour and practical reason, which is slightly more popular in psychological literature, perhaps to emphasize the deliberative character of intelligent behaviour particular to humans and some animals. Consequently problem solving, decision-making, or choice, as economists in turn prefer to term it, will be used as synonyms hereof.

Process We consider useful to introduce a primordial ontological category *process*, as this term will reappear across the thesis in different contexts. Let us define the process as a series of events that transform certain qualitative endowment (input) at certain time t_0 into a different qualitative endowment at t_1, where $t_0 \neq t_1$. Process is one of the fundamental, and for process philosophers the very fundamental ontological category. Prominent process philosophers include Heraclitus, Leibniz, Bergson, Peirce, William James but foremost Whitehead, who believed that in a dynamically changing

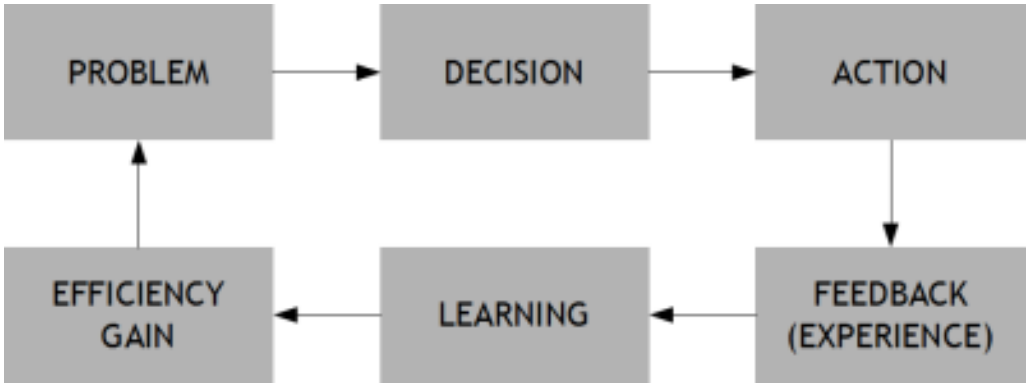

Figure 2.1: Purposive behaviour efficiency improvement cycle

universe process is a principal category of ontological description (Rescher, 2009). Interestingly, recent insights from neurological studies of human brain suggest that *self*, one of the primary ontological categories is indeed a process that is present at all times when one is conscious, rather than a thing (Damasio, 2010).

Let the purposive process be the process that is a result of an agent voluntary action aimed at achieving a goal, transforming the given state of universe into the desired output state, acting upon agents experience. [1]

Importantly the resulting output state does not need to be equal to the goal state as agent action may not be efficient. Therefore an instance of the purposive process can be represented as a tuple:

$$S =< U_0, U_g, U_1, E_0, E_1 >$$

where U_0 is the input state, U_g is the goal state, U_1 is the output state and E_0 is the initial experience the agent acts upon and E_1 is the output experience of an agent. Additionally ΔU can be defined as the gap between input state and goal state, i.e. the problem. The instance of the purposive process may be referred to as *event*. The purposive process is recurrent as it loops in constant efficiency cycles due to the fact that: (i) goals are not achieved in every instance (there exists $\Delta U \neq 0$), (ii) goal state U_g may change from one instance to another, (iii) agent never stops acting, in particular if one goal is attained a different U_g is set. The purposive process efficiency improvement cycle is graphically illustrated by below figure:

Important element of the process depicted above is *learning*. Learning is the change in agent's response to problem manifested correspondingly in subsequent decisions and actions that results from the gained experience ΔE in past instances of purposive behaviour, and is an indispensable condition for efficiency improvement.

[1] Importantly the above definition of purposive behaviour holds both for individual agent and groups of agents. The different is that we will have group decisions, collective goals and collective experience and action as arguments instead. For the purpose of current discussion let assume that an agent can be replaced in a purposive process loop by a collective entity, such as an organization.

As mentioned earlier, in line with Whitehead's process ontology, it results worthy to look not only at action and learning which are naturally dynamic and time-bound but also at other, seemingly more static, phenomena such as knowledge and mind also as processes. Contemporary philosophers and neurobiologists talking about mind and consciousness often refer to this ontological category for putting forward their theories on these complex phenomena. Searle for instance puts mental processes in the center of human sciences. As particle physics is fundamental for all natural sciences the study of human mind is fundamental for social sciences. Whatever happens in the social sphere is dependant on human mind, the study of which is absolutely vital for deeper understanding of economic and social behaviour of human beings. The choice, being the result of an individual exercising one's free will is the basic starting point for analysing this behaviour, source of all which are mental process.

This brings us to the next set of terms to be introduced: *knowledge*, *mind*, *affect* and *experience*, and later terms related to subjective dimension of experience: affect, emotion and feeling.

Knowledge *Knowledge*, *experience* and *experience* (KR) will be thoroughly discussed from the philosophical angle at the beginning of chapter 3 as these are critical terms for the viewpoint of the main goal of the thesis, here however we will focus on definitions of the terms as will be ultimately used through the reminding part of the text so as to provide terminological guideline for the reader.

There are likely as many definitions of *knowledge* as there are books on the subject. It is true that it is difficult to construct an all embracing definition of the phenomenon but at the same time the term is widely used and often refereed to being in the top 1% of most looked-up entries in Merriam-Webster on-line dictionary. The definition provided by the dictionary itself, which given the procedures of maintaining the entries in the dictionary well corresponds to the daily usage of the term, distinguishes 5 contemporary meanings of the term: (ia) the fact or condition of knowing, (ib) acquaintance with or understanding of a science, art, or technique; (iia) the fact or condition of being aware of something, (iib) the range of one's information or understanding; (iii) the circumstance or condition of apprehending truth or fact through reasoning; (iv) the fact or condition of having information or of being learned; (v) the sum of what is known: the body of truth, information, and principles acquired by humankind. (*knowledge*)

This definition provides a good overview of how people usually understand the term knowledge in everyday language. Still, if we are talking about knowledge and its representation in artificial systems such a rough definition will not suffice. Besides, apart from the first understanding of the term as "the condition of knowing" other ways of understanding it especially "the sum of what is known" or "apprehending truth or fact" may be misleading for our purposes and shall be abandoned. Instead, let us step back and look at knowledge as the product of a *mind*. Searle (Searle, 2004,

p. 8) proposes to look at the mind as the central topic in philosophy and approach all other contemporary philosophical key topics, such as the nature of language and meaning, the nature of society and the nature of *knowledge* as special cases of the more general characteristics of human mind. Knowledge is evidently both a substratum and a product of the mind, which well corresponds to the idea that both mind and knowledge are processes rather than entities. However in order to be able to harness this process for pragmatic purposes we need to realise which are the instances of it. The instances of the mental process are evidently states of mind, given states of brain, central nervous system and the entire body of the mindful subject. It is necessary to realise that knowledge cannot exist neither can have any significance until it appears in a mind. A counterargument can be proposed that a computer program that has some procedural knowledge encoded can have an effect on reality without appearing in any mind. This is true, however procedural knowledge of a type how to process one input state to another will not be regarded as knowledge in the main meaning of the term we are discussing here. The fact that knowledge is represented makes a difference, and knowledge represented in an artificial system cannot be by nature the same as knowledge as a state of mind.

Talking about the *mind* it is necessary to make reference to related terms *brain* and *body*. The main-body problem is the classical philosophical problem source of which is the dualistic view introduced by Descartes on that the realms of though and physical reality are two principally different entities. Many contemporary philosophers, on which we will elaborate in Chapter 3, reject dualism by which the classical mind-body problem ceases to exists, as there is unity between mind and body, according some, or there is nothing but physical brain, as claim others. We will favour the non-dualistic approach at many levels across the thesis by: (i) accepting mind-body dualism rejection, but through mind-body integration rather than reductionistic elimination of mind, (ii) accepting mind-brain unity, and (iii) questioning the knowledge-affect dualism. Talking about mind-brain unity let us note that a few distinguished neurobiologists, including Jaak Panksepp, use this compound as a single word *mindbrain* (Panksepp, 1998) to acknowledge that indeed mind and brain are one, as there is nothing in the mind that would not exists in the brain and vice versa. Importantly *brain* in this case may be understood more widely as *central nervous system (CNS)* that anatomically is composed the brain and the spinal cord which together build up the majority of the nervous system.

As anathomy of CNS is not central for us here whenever we use the term brain or CNS hereof we shall mean the biological implementation of mind, whichever the anatomical boundaries of this implementation in nervous system are, leaving this debate to other fields, or we will simply use the term mindbrain after Panksepp.

Experience *Experience* is a term and concept that is widely used across many disciplines of science and in everyday language. Interestingly, a query ”experience” in

Encyclopaedia Britannica On-line (*experience*) results in 5,936 hits but there is no entry dedicated to the term, similarly in the MIT encyclopaedia of cognitive sciences the term is used excessively across entries still it is not indexed (Wilson and Keil, 2001). Historically *experience* was primarily dealt with in philosophy, epistemology in particular, and psychology. Currently the interdisciplinary cognitive sciences: cognitive psychology, philosophy of mind and neuroscience delve into experience as a cognitive and brain phenomena. As in case of *knowledge* no common terminological consensus has been worked out in this respect. Apparently talking about experience in scientific terms appears to be a challenge, nevertheless some important endeavours have been made and will be hereby presented.

To start with, it is important to distinguish three most common meanings of the word *experience* used in different contexts. Firstly, (i) experience can be associated with procedural knowledge or skills (know-how), secondly (ii) experience can mean the body of knowledge about the environment that is gained by participating or assisting different events over time, finally (iii) it can be understood, in line with the psychological definition by Eysenck as appreciation of stimulus event or the knowledge resulting from this (Eysenck, Arnold, and Meili, 1972). In this thesis we are primarily considering experience in the contexts as outlined in (ii) and (iii), the latter in particular.

The closest to our understanding of experience is the above quoted Eysenckian definition of experience: subjective conscious appreciation of stimulus event or the knowledge resulting from this, albeit we would rather say 'and' instead of 'or' as to underlined the fact that experience is composed of both a subjective and objective component which are inseparable, to which we will pay much more attention in Chapter 3. Although the above-mentioned definition is variously rephrased, all of the definitions found in the literature point at the following particularities of the phenomena:

- experience is subjective, as it is born in the mind of an individual, so although we can talk about group experience (Littlepage, Robison, and Reddington, 1997) it will always be as derivatives of individual experience;

- experience emerges as a result of changes in the environment, as experience is triggered by events involving the subject as an observer or active participant;

- experience is a result of complex cognitive process, as experience is primarily shaped by perceptions, but also emotions, previous experiences, and interpretation assigned to it by the human mind (Johnson, Verfaellie, and Dunlosky, 2008);

- experience is memorized, i.e. it is registered by memory (Reyna, 1995), and as such it is translated into new knowledge, thus it plays a key role in learning (Morgan, 2002), consequently experience is accumulated and constantly

transformed by human mind in a very dynamic process influenced by memory volatility;

- experience shapes behaviour by building an internal replica of the external world allowing the subject to orient oneself and adapt its behaviour to the given reality (Lehar, 2006) in the pursuit of one's goals;

- experience defines existence, a lesson taken from Nagel who in his famous essay "What it is like to be a bat?" noticed that conscious experience defines act of being, underlying thus the very subjective character of experience and its fundamental role for existence; mind is that which thinks and experiences, thus thinking and experiencing defines a minded creature, including a human being (Graham, 1998);

- experience is a product of both conscious and unconscious mind processes, as Lashley provocatively stated "no activity of the mind is ever conscious" (Lashley, 1958), as all that is conscious in human mind is build on the inaccessible unconscious information processing in the brain, and increasingly contemporary brain research provides more evidence for that most of what the brain does is outside conscious awareness, Eagleman for instance concludes that consciousness is "the smallest player in the operations of the brain" (Eagleman, 2011, p. 5).

The above points provide a fair overview of how phenomenon of experience is approached by psychology, cognitive science and AI. We will not take up any discussion leading to the refinement of the concept now, leaving it for Chapter 3, instead we will rest upon the definition proposed in (Kaczmarek and Ryżko, 2009) concluding that *experience* is mind's appreciations of stimulus events, accumulated in memory over time that embrace but at the same time transgress knowledge. In slightly simpler terms experience could be seen as a sequence of remembered states of mind shaped in cognitive process (process of brain activity determined by internal and external stimuli) the purpose of which is to allow the subject to pursue one's goals. We will refine and build on this provisional definition later in chapter 3.

Before continuing with presenting further terms related to subjective dimension of experience let us reflect briefly on experience quantification and measurement. In view of the above terminological discussion accumulating experience stands as a complex psycho-cognitive process of particularities that make experience difficult to measure, quantify, model and optimise. According to most of psychological theories experience which is critical for human behaviour is: (i) complex, (ii) subjective, (iii) dependant on the unconscious, (iv) dynamic, (v) intangible.

Although visibly experience is a fuzzy concept it is evident that experiences can be assessed in a qualitative or event quantitative way. Experience can be positive, when results in feeling of pleasure, or negative when results in lack of pleasure or pain. Furthermore experiences could be ordered according individual preference as

one can say that experience x is better, worse or equals to another experience y. Naturally, as experience itself is intangible this will be ascribed to a given stimulative event (whereas physical, sensual, emotional or mental) causing experience x and y. Evidently, as experience is subjective the same stimulative event can result in experiences of different quality for different subjects, or for different circumstances (time-space dependence) This makes experience assessment and comparisons highly difficult and complex, though psychological literature provides some examples of both theoretical and empirical endeavours challenging this problem. Maslow developed a notion of peak experience (Maslow, 1971), which he defined as the moment of highest happiness and related it to self-actualisation. Similarly, the optimal experience was described by a Hungarian psychologist Csikszentmihalyi, who introduced the concept of *flow* (Csikszentmihalyi, 1975), the moment of top experience when one is confronted with a demanding challenge, still attainable with one's own capacities, and deeply enjoys the moment of stretching intellectual capabilities, and thus learning and increasing self-esteem. We can therefore talk about maximisation of subjective experience.

Schmitt, considering experience in the Customer Experience Management context (Schmitt, 2003) proposes that customer experience comprises: (i) customer satisfaction, linked to functional aspects of product or services, (ii) customer emotions, linked to psychological comfort or pleasure, and (iii) social comfort achieved by social fulfilment.

Although Schmitt account of experience appears to us as simplistic and inconsistent, for as it implies unnecessarily partiality of 'components' of experience as well as introduces dichotomy between social comfort and emotions which is not justified, he managed to bring the notion of experience to attention of a wider public, business people in particular, creating demand for both scientific debate on and tools for experience quantification and representation in information systems that lead to advancements in the area (Kaczmarek and Ryżko, 2009).

Emotion, feelings, affect and qualia The discussion in Chapter 3 will lead us to the conclusion that knowledge is not equal to experience, that there is a differentiating factor, which makes experience a broader phenomena than knowledge, although readings by giants such as Aristotle and Einstein has been interpreted historically as to propose the contrary, i.e. that knowledge and experience are one. This may be the issue of terminological framing or a deeper disagreement at the level of ontological principles upon which we leave the reader to judge having gone through Chapter 3. The differencing element is the subjective component of experience, the "subjective appreciation" as Eysenck put it, or 'qualitative feeling to it' as it is called contemporaneously (Damasio, 1999).

Unveiling the conclusion of the detailed discussion in chapter 3 let us point out that the subjective component of experience is closely related to which in psychologists

label *emotions*. Since the terms related to subjective component of experience such as: affect, feeling, emotions, and qualia are used in non-psychological literature, especially information science literature, rather loosely introducing unnecessary confusion, let us briefly discuss these terms here. Part of this discussion will echo in Chapter 3.

Modern psychology defines emotion as negative or positive reaction to a perceived or recalled object, event or circumstance accompanied by a subjective feeling (Damasio, 1999). The evolution of emotion theories from classical Cannon-Bard theory (Cannon, 1927), psychological account of emotions by Lazarus (Lazarus, 1991), under which emotion is a function of subject's appraisal of the situation, cognitivist approach (Oakley, 1992; Solomon, 1973; Neu, 2000) popular especially in the second half of XX century which proposed that emotion is a cognitive phenomena and occurs in consequence of cognitive processes, to recent ground-breaking findings by LeDoux (LeDoux, 2000) which suggest that emotions are result of both physiological reaction of brain and body, and/or mental interpretations related to a given situation involving different brain subsystems for different types of emotions, clearly shows that emotions should be regarded holistically as neuro-cognitive and mental phenomena. Early XX-century emotion theories regarded cognitive and emotional processes as separate yet dependant entities and concentrated primarily on the problem of sequentiality of emotion and cognition, that is trying to address the question: which comes first emotion or cognition and how these two interrelate. Later in the era of domination of reductionism emotions where relegated to the mere resultant by product of cognitive appreciation of external stimuli. Under contemporary accounts of emotion which depart from emotion-cognition dualism this problem has been replaced with the search for the nature of the subjectivity of consciousness that build up the feelings of emotion. Importantly this led to distinction between emotions and feelings, where the former are defined as hard-wired programmes that are triggered by a stimulus and comprise a set of relatively fixed bodily responses that are only modulated by consciousness to certain extend, meanwhile the latter correspond to conscious subjective appreciation of the bodily state provoked by emotion programme. We owe the view on emotions as 'affect programmes" largely to Ekman who inspired by Darwin investigated emotional bodily expressions across cultures and evidenced emotions as complex responses typical of all human beings across races and populations, which are controlled by mechanisms operating below the level of consciousness, identifying furthermore a concrete set of these programmes which included: happiness, sadness, fear, anger, surprise, and disgust (Ekman, Friesen, and Ellsworth, 1982; Ekman, 1989; Ekman, 1993). Contemporaneously this approach is continued and expanded by neuroscientists such as LeDoux (LeDoux, 1996) and Panksepp (Panksepp, 1998) that investigate emotions at the level of neuronal circuits in the mammalian brain. *Feelings* on the other hand both in Damasio's and Panksepp's account are different in nature and correspond to the subjective appreciation of bodily states coming from within the body as well as from outside via sensory gates. The particular characteristics of such

defined feelings are their origins in evolutionary old parts of the mammalian brain for which reasons these are called by both neuroscientists *primordial feelings*. Under both accounts primordial feelings provide basis for emergence of self. Separating emotions from feelings and specifying emotions as programmed responses created a vacancy for a more general term that would embrace subjectivity of experience in all forms: emotions, feelings, incl. feelings of emotions. The term that has filled in the gap is *affect*.

For introducing the term *affect* let us start with quoting Dr. Jaak Panksepp, author of a widely respected book "Affective Neuroscience" and one of the pioneer of the field gaining more and more attention recently.

> "I chose that word for coining the term 'affective neuroscience' to highlight that neuroscience has yet to deal with the nature of our positive and negative experiences-the many ways we can feel about things instinctively. Why does sugar taste wonderful on our tongue, but not a cat's?" (Panksepp and Campbell, 2010)

Miriam-Webster defines affect as "the conscious subjective aspect of an emotion considered apart from bodily changes, also: a set of observable manifestations of a subjectively experienced emotion" (*affect*). Other popular dictionaries like Oxford emphasize the more etymological meaning of the term as "emotion or desire as influencing behaviour". Although these are common meanings of the words as used in everyday English they very aptly reflect their usage in scientific literature. In brief affective is a fuzzy term under which anything that cannot be better described and is related to falls within.

Cognitive philosophers of mind include affection in the so called trilogy of mind, next to cognition and conation (Hilgard, 1980). This places affect in the very centre of existence and behaviour. Behaving agents perceive environment, this perception has a qualitative subjective, i.e. affective, value which is complemented by their natural tendency to act, i.e. conation including rationality, purposiveness of action and motivation, we could also refer to it as *the executive* component of mind. We would later try to show that there is rather unity of intentional contents and their affective complement, unified with the rules of behaviour encapsulated in the relationship between the cognitive and the affective. This mapping of the cognitive and the affective onto action forms the executive, which is not by nature a separate entity however it can be separated for the purpose of theoretical deliberations and rationality modelling.

Given the above, as well as the fact that affect is subjective and is characterised by valence allows us to risk to state that there is a direct link between affect and quality of experience. Affect shaped by feelings of emotions could be simply seen as an integral component of experience that complements knowledge and determines experiential valence (positive-neutral-negative) and intensity (high-low). In this

perspective, the problem of sequentiality of emotion and cognition, in other words what comes first the emotion or cognition and how these two interact becomes irrelevant. What gains relevance in turn is how affect-valenced experience is shaped and how it maps onto action. Another important arising question is how this looks in the dynamic environment, what role in this mapping has memory to play, as plausibly current experience depends on previous experiences. Furthermore own experience provides the basic material for the mind projections into future, which anticipate future action.

Having considered experience in this perspective and bearing in mind the ultimate goal to be able to quantify experience for the purpose of its representation in information systems it is inevitable to recall a vivid debate which started in 1980s on *qualia* (singular: *quale*). Qualia are defined as experimental properties of sensations, feelings, perceptions, thoughts, desires, etc., in other words qualia include what it is like to have experiential mental states or, yet in other words qualia are ways things seem to us (Dennett, 1988). To explain the term by example let us take for instance the very particular, subjective and personal auditory quality of hearing to the church bell ringing at a given moment, this quality would be a auditory quale of the subject listening to the ringing bells. Such a definition of qualia makes them, in view of what we have said earlier equal to subjective content of experience. For this reason the debate on qualia is highly relevant to this thesis and will be reviewed and the concept of qualia will be revisited by us in chapter 3. This is particularly important as the term has raised controversies ever since its first appearance and some philosophers reject they important role presenting very convincing arguments, which cannot be left unanswered.

2.3 Terms related to information systems

The discussions over feelings, emotions and experience in this thesis are carried on with the single objective, which is representing these subjective phenomena in information systems. Let us in this sub chapter briefly introduce the key terms in this discussion so as terminological clarity and consistency could be kept from as early stage as possible. These terms include some fundamental terms for knowledge engineering and information science such as representation and processing of knowledge, information systems, information retrieval systems, representation of emotions. Finally we will make a brief note on affective computing, a relatively young yet quickly growing branch of AI. Knowledge Representation (KR) is one of the main fields of artificial intelligence(AI) which objective is to provide efficient methods for symbolic representation of human knowledge in artificial, formal systems, in particular in so called intelligent or expert systems, for that consequently the represented knowledge could serve as basis for machine problem solving, i.e. decision making, and generation of new knowledge by inference from the knowledge priorly

stored within the system. AI researchers has taken the challenge of identifying efficient ways for explicit representation of knowledge in computer programmes and its algorithmic processing as a way towards designing artificial intelligence systems formalized as symbolic reasoning automaton (Van Harmelen, Lifschitz, and Porter, 2008). Muraszkiewicz (Muraszkiewicz, 2011) defines KR as a methodology of presenting the knowledge about the world along with the procedures for processing this representation, especially by means of reasoning (inference) and proposes a formal definition as the following duple:

$$\mathrm{KR} =< \mathrm{KR}_{\mathrm{Description\_Language}}, \mathrm{KR}_{\mathrm{Processing\_Mechanism\ (incl.\ Inference\_Mechanism)}} > \quad (2.1)$$

It must be underlined that the field of KR uses a much more formal definition of knowledge, which is dictated by the necessity of exactness and strong influence of the deterministic thought on the domain. First of all KR builds on more specific definitions distinguishing different kinds of knowledge. One of the most important distinctions is that between *procedural knowledge* and *propositional knowledge*, where the former captures the knowing *how* to do something, eg. driving a car, and the latter the knowing *that* some particular proposition is true, e.g. knowing that earth is a planet (Ryle, 1949). A third type of knowledge that is distinguished form the other two is that which captures knowing a place or a person (Steup, 2008), which is knowing in the sense of being acquainted with something. All the three meanings are seemingly significantly different in nature and KR primarily focuses on propositional knowledge, also referred to as declarative or semantic knowledge. This is because procedural knowledge includes, or is the basis for the other two. Procedural knowledge bridges the propositional contents of mind with behaviour. Knowing how is a combination of knowing that with goals driving purposeful behaviour, or in other words the procedural knowledge is the mapping of means to ends. With this approach procedural knowledge can be ontologically reduced, however remains a useful term for describing the mapping itself. Similarly knowing as a condition of acquaintance with x can be reduced by knowing that x exists and knowing that x has a given qualitative endowment. Consequently, in principle the concept of knowledge which is of particular interest to KR researchers, and us under this thesis, is of declarative kind, and the earlier adopted general definition of knowledge, i.e. knowledge as intentional contents of mind, complies. However knowledge engineering and KR makes it more narrow. The subject of analysis under KR is the concept of knowledge as expressed by the schema "S knows that p", where S refers to the knowing subject, and p to the proposition that has a logical value (Steup, 2008). Similarly Muraszkiewicz states that knowledge is every entity that can be expressed in language and has logical value (Muraszkiewicz, 2011). This discussion will be continued in Chapter 3. Contemporary thinking about what is knowledge and what are the contents of mind is under heavy influence from the legacy of philosophy of language. The famous Wittgensteinian claim that "The limits of my language are the limits of my world" encapsulates

knowledge in language, which KR takes for granted and on which it rests entirely. Meanwhile non-linguistic forms of intentionality of mind as well as the subjective dimension of mental states, the role of emotions in building language capacity and its underlying role for enabling mammalian brainmind with intentional capacity, has been persistently ignored by KR, which we are trying to address herewith.

Another aspect of knowledge typically ignored by KR is its relation to consciousness. An important question arises: is knowledge only that which a subject is aware of or it includes also states of mind which are not available to subject's consciousness, still have intentional contents and provide meaningful guidance to subject behaviour? Let the provisional answer here be that both conscious and unconscious intentional contents of mind qualify as knowledge, however let us make a terminological distinction so as we will refer to these two as *explicit* and *implicit knowledge* accordingly.

In above paragraphs we have used terms such as artificial intelligence, expert system and knowledge engineering that need to be explained and relations between them shall be traced.

Artificial intelligence was defined by Minsky, one of the pioneers of the field, as "the science of making machines do the things that would require intelligence if done by men" (Minsky, 1968). Cohen an Feigenbaum provide a more specific definition "AI is the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behaviour - understanding language, learning, reasoning, solving problems, and so on" (Cohen and Feigenbaum, 1982). Rutkowski shortly states that AI is about exploring human intelligence and its implementation in machines (Rutkowski, 2005). Rutkowski rightly notes that despite the fact that AI is primarily the branch of research in computer science it resonates with many other disciplines, for instance: information science(e.g. information theory, representation, storage and processing of information and knowledge), philosophy (e.g. consciousness, experience, epistemology, ethics), biology and medicine (bio-physiological aspects of human cognition, reasoning, information processing by the nervous system, particularly the brain), logic and mathematics (eg, inference, reasoning, mathematical modelling of cognitive processes and decision-making), psychology (cognition, affect, motivation) and computer science (eg, intelligent systems, machine learning, algorithms, expert systems), for which reasons it qualifies as highly interdisciplinary branch of science.

For AI the fundamental challenge is the design of methods and techniques of knowledge representation in the form of symbols that can be processed by machines, Turing machines in particular, for inferring about actions and new knowledge generation. Consequently one of the key outputs of AI are *expert systems*. Expert system is a computer program that uses knowledge and reasoning process (inference) to solve problems that require human experience (a domain expert), acquired through years of activity in the field. The basic elements of the expert system are: knowledge base

(a set of facts and rules), the reasoning machine and the user interface for data entry. *Knowledge engineering* is the field that deals with devising expert systems, and the research challenges it addresses are: sourcing, structuring and processing knowledge, design and selection of inference methods, and finally the design of user interfaces (Rutkowski, 2005, p. 9).

AI and expert systems as a field flourished and matured in the era of computer functionalism domination in philosophy of mind. In short computer functionalism is the view that brain to mind is like computer hardware to computer software. This analogy has fuelled research in the domain of knowledge engineering. Together with the predominant philosophy of language it resulted with the focus on propositional knowledge seen as contents of mind encapsulated in language. Consequently the mental processes has been considered equal to language processing, and the field has been, apart from the earlier mentioned KR, primarily about logics, logic programming and natural language processing. If the contents of mind, i.e. knowledge, are more than language can express the question arises if the classical methods and techniques of AI and knowledge engineering suffice, and this has important implications not only for experts systems but information systems at large.

Expert systems constitute a class of a more general entity: information systems (IS). According to the definition presented in the aims and scope of Elsevier's Information Systems journal *information systems* are the software and hardware systems that support data-intensive applications. A much more specific, yet technical, definition is provided by Pawlak who by an information system means a classification systems that assigns certain qualitative endowment to a set of objects, and provides a formal definition of IS as a quadruple:

$$S =< X, A, V, p), \tag{2.2}$$

where X is a finite set of objects, A is a finite set of attributes, $V = \bigcup_{a \in A} V_a$ where V_a is the set of values of attribute a, and $card(V_a) > 1$, p is a function from $X \times A$ into V (Pawlak, 1981).

Muraszkiewicz (Muraszkiewicz, 2011) rightly notices that KR and knowledge engineering lies in the intersection of AI and IS, so the KR is the area of investigation of both AI and IS. An IS as defined by Pawlak seemingly relates faintly to AI, as it presents a purely computational scientific approach to data, information and knowledge, however the link becomes much more tight if we take the perspective of *information science* that has witnessed a profound shift over past three decades from focusing solely on the technological aspects of data and information classification, storing and access to paying proper attention to the human interaction with information technology and the usability of information systems for end user. Besides the emphasis on user perspective by the information scientist that originate from librarians or documentalists has been historically strong.

So the difference in approaches between AI researchers and information scientists with librarian or documentalist background in so far as the role of information system is concerned is that for an AI scientist IS is predominantly propositional knowledge processing machine, meanwhile for an information scientist it is predominantly the source of *meaningful* information for the system user, that is information that enriches user's knowledge. This difference in perspectives on the role of information systems, is also visible in how *data*, *information*, and *knowledge* are defined and related across these disciplines.

Ingwersen proposes that:

> "The concept of information, from a perspective of information science, has to satisfy dual requirements: on the one hand information being the result of a transformation of a generator's knowledge structures (by intentionality, model of recipients' states of knowledge, and in the form of signs); on the other hand being something which, when perceived, affects and transforms the recipient's state of knowledge. Information is seen as supplementary or complementary to a conceptual system that represents the information processing system's knowledge of its world. If only the first condition is met, we are talking about potential information, i.e. data or similar entities stored in IR systems, that is of potential value to recipients (whether humans or machines)" (Ingwersen, 1992, p. 30-37).

Such account makes the boundaries between information and knowledge blurred and emphasis is put on the role of information in shaping users's knowledge, unlike it is in computer science and AI which sets more clear boundaries between which is data, which information and how these two relate to knowledge. Zins (Zins, 2007) notices that "data is commonly conceived as the raw material for information, which is commonly conceived as the raw material for knowledge". Similarly yet more specifically Muraszkiewicz (Muraszkiewicz, 2011) explains that *data* are finite sequences of symbols, e.g. 24/06/1980, referred to as data string, *information* is data with interpretation, e.g. a date, meanwhile *knowledge* is information together with relations defined on the information set, e.g. Jan's birthday date. We will later propose that in order to be able to add *experience* into this account it is necessary to further add to knowledge the subjective feeling particular to it, which together is the basis for establishing *meaning*, meaning that goes beyond semantics.

To complete the picture let us quote Ingwersen who has cast interesting light on the relation of information science, which has emerged from library science and the study of scientific information, its documentation and the processes involved in scientific communication, covering with time also the wider aspects of scientific investigation of the processes of generation, representation, management, retrieval and use of information, with other tightly related disciplines: information theory and computer science:

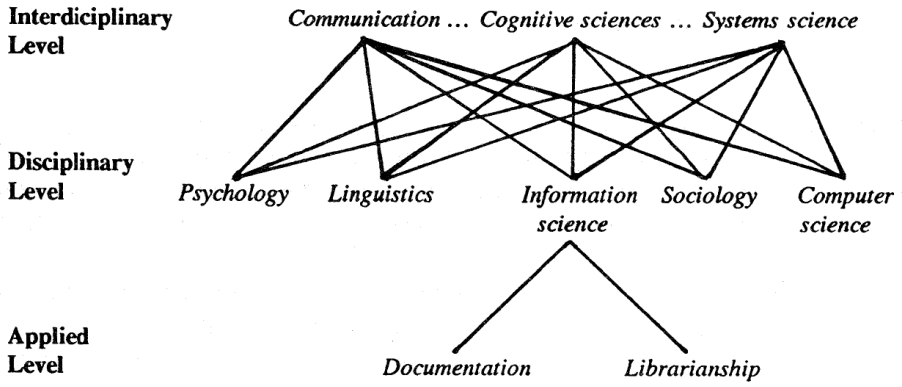

Figure 2.2: Information science viewed as one of the several sciences of information (Ingwersen, 1992)

> "The problems for information science with respect to its borderlines with other disciplines are mainly found at interdisciplinary level, less often at the disciplinary level. A core dimension noticed by other fields, is that information science actually is the one which studies large text entities containing preserved knowledge – with more interest in solving theoretical and practical problems of its organisation and representation in systems for later retrieval and use on demand, than in the technology itself; the latter being the means to the former. Consequently, important areas of common interest between information science and other disciplines may develop. One may state that its applied level contributes to its recognition."

Furthermore Ingwersen tracks the relationships of information science with other disciplines

The above quoted paragraph by Ingwersen aptly touches the very essence of information science and the ultimate purpose of information systems, which is to provide meaningful answers to problems reported by the user based on the body of knowledge stored within and available for retrieval at any time.

Before elaborating on this point further let us briefly introduce the underlying concept here: *information retrieval* (IR). IR was one of the key pillars of information science and *information retrieval systems* (IRS) its key applications. Sosińska-Kalata (Sosińska-Kalata, 1999) defines IRS as communication channel, which provides for the transmission of the pieces of knowledge fragments between the structures of public knowledge represented within the system and the structures of private knowledge represented in minds of users manifested in queries formulated by them and entered in the system. So information retrieval systems is special type of information system that provides the interface between public knowledge and private knowledge. Thus,

in IRS informations science as viewed by Ingwersen fulfils its mission: provides meaningful information that enriches user's knowledge.

Historically this could have little to do with AI. However with progress of AI techniques the concept of intelligent or semantic information retrieval could emerge as a natural evolution of traditional information retrieval systems of which early progenitor is a library catalogue. These are intelligent IRS where information science meets AI. Belkin (Belkin, 1996) doubts if intelligent IR can be achieved by sole merging of IR with AI, and bets on the human factor for providing intelligence to the IRS. He also points to the fact that *intelligent* in the context of IR is a vague concept and refers to Croft (Croft, 1987) who identifies intelligent IR with merely a *good* IR, i.e. that results in effective performance, noting that it inappropriate to ascribe intelligence to a computer program. Although there is merit in this claim we think it is possible to define IRS as the system capable of solving an ill-structured problem defined by a user and related to the domain of IR.

Let it be noted that in the same way as today the role of a traditional library evolves, from a place for storing and providing access to books and other printed and digital carriers of information to a place where one can find support in solving problems, not only information related problems but problems of various types: business, legal, etc., still primarily by ensuring easy access to quality, i.e. meaningful, information (Frey, 2006) the evolution of information systems is directed towards supporting informed intelligent solving of users' problems, formulated in natural language, rather than passive return of information based on a formalized query. This is in particular a topical issue in the times of information overload (Toffler, 1984). Meanwhile consulting a classical information retrieval system can be compared to searching a library catalogue, consulting intelligent or semantic information retrieval system could be compared to asking a librarian for help in solving some problem related to information search.

Contemporary most powerful information retrieval systems: Internet search engines pursue predominantly one ultimate development goal: how to effectively and intelligently address users problems stated as questions formulated in natural language entered into search bar, rather then returning simple results of text queries. Semantic browsing, intelligent search input bars have become the topics on which information scientists focus their research on. Providing meaningful information in long term cannot ignore the affective dimension of knowledge and we foresee increased interest of IRS in affective aspects of knowledge and how IRS as "communication channels" can serve the purpose of effective transmission of experience, i.e. both knowledge and its subjective, affect-determined quality. This is the vision that goes beyond semantic web, it is about *affective web*, where both knowledge and experience can be shared.

This brings us to the final terms we wish to introduce in this chapter: *affect representation*, *affective computing*, and finally *affective information retrieval system*.

Given what has been said in previous paragraph we shall start with affective information retrieval system which would simply be the interface between public, i.e. shared, experience and private experience. This interface should be capable of transferring meaningful information that enriches user's experience not only knowledge. In order to be able to work on this kind of systems it is necessary first to investigate the possibility of affect representation in information systems to which we dedicate large parts of this thesis. Here let us just state that affect representation seeks for methods and techniques of non-linguistic representation of affective subjective states enriching knowledge and ways for its processing and emulation. There is an emerging interdisciplinary branch of computer science that deals with this matter: affective computing (Picard, 1997; Scherer, Bänziger, and Roesch, 2010) which contributes to the larger field of affective sciences (Davidson, Scherer, and Goldsmith, 2003), which for the moment seems not to include affective information science to which we would like to contribute, albeit there are already publications in the area of information science that deal with the importance of affect to information behaviour search and theory identifying new affective paradigm within (Nahl and Bilal, 2007).

2.4 Conclusion

In this chapter we have introduced the key terms relevant to the subject matter of the thesis as well as we have put the problems discussed herewith in the multidisciplinary context. The key terms has been introduced with emphasis on presenting differences in accounts of the concepts behind these terms between different fields that overlap with the subject matter of this thesis. Preliminary own definitions of knowledge, experience and affective information retrieval has been proposed. The terminological discussion presented in this chapter will echo in following chapters as it is closely related to the objective of this mostly theoretical dissertation.

Chapter 3

Towards affective theory of experience

The objective of this chapter is to first elucidate the seemingly fuzzy term: *experience* and then synthesize with our own account of experience. The analysis of the concept will be carried out mainly form the point of view of philosophy of mind, cognitive neuroscience, psychology and information science as to provide for a possibly wide yet contemporary perspective on the phenomenon.

To achieve the goal of this chapter we will start with contrasting knowledge with experience, which shall lead us to the preliminary formulation of our revised account of experience and its relation to knowledge. In chapter two, we will look into the nature of experience by analysing qualities of consciousness that are inherent in experience. The most critical properties of experience that are subjectivity and temporarily will be discussed in separate sections. The discussion on subjective character of experience will echo the debate on qualia which are in philosophy of mind considered experience quanta. Then the affective dimension of experience will be discussed as it touches upon the pivotal element in our account for which reason an entire section is devoted to this topic. Finally, links between experience and rational agency will be established.

This is done with the purpose to propose a formal definition of experience and a framework for representing experience and emulating natural agency in information systems, which will be the main subject of the next chapter.

3.1 Is knowledge equal to experience?

In this section we will try to answer the question from the title which in principle is how knowledge relates to the subjective dimension of experience. Unlike the other questions that have been discussed extensively by philosophers since ancient times the question about subjective components of experience and its relation to knowledge has not been the central point of focus of philosophical debate, at least not so framed.

Only fairly recently contemporary philosophy of mind, cognitive science and AI research in the XX century focused on emotion and affect as relevant components of knowledge and behaviour and has tried to track relations between these concepts and experience. This question is of key importance for the thesis proposition, as we have claimed that what differentiates knowledge and experience is exactly what escapes the attention of mainstream knowledge representation theories, that is: the subjective dimension of experience, still being an inseparable element of knowledge in the general meaning of the term, as a basis for an agent getting self-oriented in the world and taking action.

This section will discuss the nature of knowledge and experience with the purpose of separating the differentiating component of our interest. To start with let come back again for a wile to the definitions of both terms.

The questions what is *knowledge* is philosophical, epistemological to be more precise, in nature. Philosophers have debated over millennia on a few fundamental questions related to human knowledge and cognition which include: (i) what is knowledge in general sense (nature, ontology of knowledge), (ii) what does it mean that someone knows something (meaning of knowledge), (iii) how knowledge is created (possibility of knowledge and methods of cognition), (iv) how can we be sure that what we know is true (the problem of truth), (v) what can we know and what we can never be sure about (limits of knowledge), (vi) what is the difference between to know something and to experience something if there is any (subjective dimension of knowledge). There is not enough space for a solid discussion on all of these topic in this thesis, for which reason it is necessary to ration the scope of focus only to those that are most relevant to its main theme, i.e. the first (i) and the last one (vi), with particular emphasis on the sixth epistemological question from the above list, i.e. what is the difference, if any, between knowledge and experience.

A few definitions of knowledge has been quoted and discussed in chapter 2 where we concluded to rest on the following one, which is an adaptation of a definition proposed by Muraszkiewicz (Muraszkiewicz, 2011), largely influenced by intentional epistemology and linguistic philosophical tradition:

> "Knowledge are all the intentional states of minded-selves that can be represented in language (natural or artificial) or other non-linguistic intentionality expressions that have a logical value determined by its correspondence to the real world or possible abstract worlds."

In chapter 2 we ended up with a definition of experience as proposed by Kaczmarek and Ryżko (Kaczmarek and Ryżko, 2009) as remembered states of mind resulting from appreciation of stimulus events that determine generically any human behaviour. Let us take this as a starting point and work o this definition further.

The elements of this definition that remain unclear are: (i) what is the relation between the stimulative event and the state of mind?; (ii) how experience is remem-

bered, as a set of past discrete events or is somehow amalgamated into one, unified sort of total experience?; (iii) how conscious and unconscious shape experience and which dimension is covered by the proposed definition?; (iv) what can we say about the subjective quality of experiences?; (v) what is the relationship between knowledge and experience. In this section we will concentrate on the last element meanwhile the rest will be discussed throughout the remaining body of the chapter.

Einstein said that "The only source of knowledge is experience" (Einstein, Podolsky, Rosen, et al., 1935), which suggests that knowledge can be considered as instances of human experience. Another classic quotation attributed to Einstein is "Knowledge is experience. Everything else is just information" (Zeleny, 2002). For Einstein there was a clear distinction between knowledge and information but knowledge and experience was one. The important question here is whether we can reverse this statement and say that any instance of experience is knowledge? To challenge this question let us come back to the considerations on what is knowledge in general, in simplified classic terms whether knowledge is *episteme*, *doxa* or *endoxa*? Importantly however the solution to this problem is *not relevant to* the main argument of this thesis. Regardless the approach taken on the knowledge-truth problem, or regardless the eventual definition of knowledge, we can rest on the assertion that knowledge always would inherit the qualities of human consciousness, as an experiencing conscious human mind is the only source of knowledge, but also its ultimate end-user.

Importantly however the mainstream definitions of knowledge proposed by knowledge representation theorists (see Chapter 2) rely on language, for which reason they cannot encompass all the qualities of conscious experience. Alfred North Whitehead rightly noticed that "An enormous part of our mature experience cannot be expressed in words" (Whitehead, Griffin, and Sherburne, 1979) And this is not because some words are still to be created to name the knowledge that is yet to be uncovered, but because of immanent structural limitations of language as language is not a direct but indirect mean of human cognition [1].

Which is that Whitehead had in mind while saying "enormous part" of experience that escapes the expressive capacities of language? The two following sections will consider this question in detail, now let us preliminarily propose that this missing part includes at least three elements:

1. Subjective qualitative feeling inherited from the consciousness of minded-self;
2. The unconscious component of experience, which will be left in further discussion as we have rationed the scope of the argument to the conscious experience;
3. Emotions.

As noted in chapter 2 the boundary between emotions and feelings is blurred, however later we will try to explain this relationship in more depth. At this stage it is

[1] See subsection 3.3.2 for further discussion

important to highlight that they are separate phenomena, furthermore emotions are much better studied and understood than consciousness, and theories of consciousness that embrace both emotionality and the subjective feelings of consciousness are relatively recent and not fully established yet.

Let us conclude this section with a hypothesis that the definitions of knowledge that impose on the scope of knowledge the limits of language are not capable of embracing the concept of knowledge in broad terms as nurtured by Einstain and Whitehead, but not in the ontological sense but cognitive sense. The missing component is mostly the subjective quality of consciousness and emotions. Importantly, as we will later try to argue, the omission of this component - the inner, intentional qualitative dimension of experience - makes the knowledge representation lame and imposes limits on real life applications, this is because ultimately it is the experience that determines the rational agent's purposeful behaviour, and because any minded self can have access only to this form of knowledge.

3.2 Consciousness as the playground for knowledge and experience

Philosophy of mind that has become the central theme of contemporary philosophy has replaced to certain degree the traditional ontological and epistemological questions with debate on the nature of consciousness.

> "(...) consciousness is the condition that makes it possible for anything at all to matter to anybody. Only to conscious agents can there ever be a question of anything mattering or having any importance at all." (Searle, Dennett, and Chalmers, 1997, p. xiv).

As Searle proposes above anything that happens in mind, including knowledge and subjective experiences depends on consciousness and inherits qualities from the more general phenomena that consciousness appears to be. In this section we will provide and overview of the contemporary account of consciousness with the purpose to identify its qualities that shape the nature of conscious experience.

3.2.1 Types and states of conscious experience

Before we start to enumerate the qualities of consciousness we shall make a short note on the distinction between conscious and unconscious experiencing. It is recognized that as experience is a product of activity of human mind it depends on both conscious and unconscious mind processes. Karl Lashley provocatively states that "no activity of the mind is ever conscious", as all that is conscious in human mind is built on the inaccessible unconscious information processing in the brain. Despite

the acknowledged significance of unconscious brain processes to both experience and behaviour we have regretfully decided to exclude the unconsciousness as the factor influencing the quality of conscious experience. Only in so far as it manifests itself in emotional processes that are known and described by neuroscience literature shall we try to include unconsciousness in the scope of discussion. The remaining part will be the subject of further work.

What is consciousness? Apparently this remains one of the big masteries that science has not managed to explain yet. However advancements in brain studying methods and technologies and significant popularity of the topic approached from many disciplines: philosophy, cognitive and clinical psychology, psychiatry, neuroscience, information science, artificial intelligence in particular, brought huge amount of literature on it, and which is more important, resulted in some early progress towards disentangling this mystery.

The modern study of consciousness involves a couple of underlying problems: (i) what we mean when we say consciousness, what are the kinds of consciousness we can experience, what does it mean to be conscious (concepts of consciousness) (ii) what are the qualities of consciousness, how does it feel like to be conscious (descriptive character of consciousness); (iii) how consciousness is happening, in particular how it is created by the brain (ontology of consciousness); (iv) how consciousness relates to and impacts behaviour of all sorts (function of consciousness). This questions will be looked at in this section with the objective to understand how consciousness maps onto knowledge and determines purposeful behaviour. There are also other fundamental questions related to consciousness that are of metaphysical character, referring to the problems of existence, spirituality, art and religion, which we will not discuss.

The principal trouble with consciousness is that it is purely subjective and we cannot have a direct insight into someone else's consciousness, so we can only study it via indirect methods like observation and intra-subjective comparisons. As for direct methods the only tool we have is introspection in which case we are limited to investigating our own consciousness. This causes a problem with defining what does it mean to be conscious in objective linguistic terms. Therefore it is important to make a few terminological assumptions which will make further discussion of this topic easier.

By saying that A is conscious we may mean a wide range of things, starting with any mental-like process, or a very broad understanding of consciousness as capability of sensing to the much narrower conception of consciousness that is a mental state of an awake human being. Armstrong suggested there are three main types of consciousness: *minimal*, *perceptual*, and *introspective*(Armstrong, 1979). The minimal is when any mental activity occurs, the perceptual when a subject is capable of sensing still may not have a sense of self, is not self-aware, finally, the introspective consciousness is as perception-like awareness of the subject's own current mental states

and activities, which is recurrent, i.e. involves being introspectively conscious of the introspective consciousness itself. Introspective consciousness, concludes Armstrong, is a necessary condition for the presence of awareness of a self and the past history of that self.

As talking about consciousness in interpersonal terms can currently only be done on the basis of intuition, our intuition dictates that the nature of consciousness is much more heterogeneous than Armstrong proposes. This heterogeneity starts already at the level of members of the same species, like human beings (for instance, between sexes, cultural communities, people with different mental capacities, people at different stage of mental development, but we will risk to say that likely even between identical twins), and gets bigger and bigger as we go down the evolutionary ladder comparing different species of primates, mammals, reverberates and finally the simplest forms of life: single cell organisms. So types of consciousness could be pointed on a continuum from simplest sensing capabilities of most primitive single cell organisms to different types of human wakefulness, or likely beyond that to some higher states of consciousness evoked by drugs, trances, meditation, discussion of which falls beyond the scope of this thesis. Still there are important landmarks on this continuum which can be tested with behavioural experiments or brain activity measurements. The important landmark for us is the consciousness of a healthy adult human being that can have the following 3 modes: (i) wakefulness, (ii) dreaming, (iii) deep dreamless sleep.

Having said that we must streamline that unconscious mental processes are those that are never accessible to self in all above mentioned modes of normal consciousness, not the processes that are going on in the brain when one is not awake, i.e. sleeping or in a sort of coma.

There are many accounts of what a mental conscious state is and what is its origin. We will try to present selected, to us most convincing accounts originating from philosophy of mind and neuro science.

Van Gulick (Van Gulick, 2011) identifies six main possibilities discussed in the contemporary philosophy of mind: (i) *consciousness as meta-intentionality*, which requires the consciousness be mental states that are themselves about mental states; (ii) *consciousness as qualitative feeling*, which emphasises the importance of the existence of qualitative or experiential properties of mental states, i.e. qualia (see the final subsection for details); (iii) *phenomenal consciousness*, which concentrates on phenomenal structure of consciousness, its spatial, temporal and conceptual organization of relationship between agent and external environment (so this approach contains the intentionalist view on consciousness); (iv) *first-personal consciousness* that stems from the classical article by Nagel (Nagel, 1974) in which he proposes that for an organism to have conscious experience it is necessary that there is "something it is like to be that organism", therefore it requires I and I-am-ness; (v) *access consciousness*, conceptualized by Block (Block, 1995), which sees the availability of

the informational content of experience for use and guidance by the organism as a decisive factor constituting conscious mental state; (vi) *narrative consciousness* which accounts for the consciousness that can be reduced to the stream of narrative episodes experienced from the perspective of an agent, either natural or artificial, which account is strongly supported by Daniel Dennett (Dennett, 1991) who questions the existence of qualia and views unity of consciousness as illusionary.

Contemporary philosophers of mind polarize into two main groups one composed of those that still accept Cartesian dualism (e.g. Chalmers, Penrose) and the other the supporters of materialistic, reductionistic account of consciousness as illusionary, non-existent phenomenon that is nothing more but collection of lower level processes in the brain.

The former group is led by Dennett. Dennett's account is particular because it assumes that there is no consciousness that matters, the consciousness is illusive, irrelevant phenomena crafted by the brain that builds the illusion up from sensational bits and pieces, meanwhile all the processes in the brain can be decomposed to purely physical processes and entities. He favours the view on consciousness as a *stream* of tiny building blocks of mental states that build up the illusive feeling of unified field of consciousness (Dennett, 1991). This account is intriguing as some convincing parallels can be drawn. As for instance we can easily enjoy the continuity and flawlessness of a piece of music encoded in .mp3 audio file formats that deceives our ear with a lowered bitrate, the organisational capacity of our brain could allow it to make up for the missing gaps between bits and pieces of intentional contents fed by our senses creating an illusion of continuity and unity of consciousness. Gestalt phenomenon is often given as another exemplification of this brain capacity.

Somewhat against these two mainstream approaches Searle provides his alternative account which he calls *biological naturalism* (Searle, Dennett, and Chalmers, 1997). Searle believes that consciousness is entirely the product of biological, physical body, mostly brain and central nervous system (rejection of dualism) that is at the same time irreducible condition which has a qualitative, subjective and unified feeling to it, qualitative in the sense that for any conscious state, there is something that it "qualitatively feels like to be in that state", and subjective in the sense that they only exist when experienced by a some sort of a subject, continuously over time. The key to understanding this view is Searle's deep critique and rejection of traditional concepts in studying nature of human mind influenced by Cartesian dualism. The problem according to Searle is that we erroneously believe that "mental" cannot be "physical" at the same time. Searle compares consciousness to other higher level processes in the body such as digestion, which despite being composed only of cells account for a process at higher level that has new properties and functions. Similarly CNS created consciousness that plays pivotal role in life management of the organism. Searlian *biological naturalism* is similar to *non-reductionist physicalism*. Searle defines consciousness in the following way:

> "Consciousness consists of inner, qualitative, subjective states and processes of sentience or awareness."

Unlike Dennett, Searle favours describing consciousness as a *field*, rather then a *stream*. He visualizes it as a plain or hyperplane that has some sensitivity potential and fluctuates as intentional contents come in and out of it.

It is still unknown how any of the three accounts of the consciousness proposed by philosophers of mind could be implemented in the brain. Still brain science provides important input to the debate although each side of the debate uses neuro-cognitive evidence pretty selectively. Importantly experienced neurologists and brain scientists propose their own accounts of the phenomena, a tiny selection of which will be presented below.

Modern science in general acknowledges that everything that is in mind has its neural representation, or at least causes in the biological brain which is nothing but a highly complex system composed of activated or inhibited neurons (Kandel, 2009). A natural intuition of many computer scientists is therefore to believe that if we could represent firings of neurons with a relevant precision we could emulate a fully functioning brain in an artificial system, this constitutes a philosophical stance known as computer functionalism, which builds on the mind-brain and computer hardware-software analogy. This stance remains highly controversial, and we share the opinion of Searle and other philosophers who believe that the fact that the brain is materialised in leaving biological tissues, not in silico, does matter (Searle, 2004). Nevertheless, the analogy is highly appealing, especially if we make use of it not to create artificial intelligence but merely represent mental phenomena in information systems for the purposes such as knowledge management, decision support, and alike.

For the above reason it is worth investigating what is the account of consciousness and self from the perspective of contemporary neuroscience. It is important to highlight that the field of neurobiology and brain science is challenged with the incredible complexity of central nervous system for which reason as a general rule the main effort under this field is invested in studying tiny little aspects or processes at a lower level of brain regions or neuronal circuits, focusing on carefully selected problems or sub-problems. The studies that would embrace higher level complexities and focus on more general phenomena and concepts such as consciousness and memory as a whole are less common and normally end up with a more or less likely hypotheses rather than sharp answers. Means of verifications of these hypotheses are not easy to come up with. So despite these are typically based on a solid scientific foundations provided by research results at the lower level the general frameworks usually are provisional, and there is virtually no complete and unquestioned theory of any general brain function. A good evidence to this claim is the recently discovered *brain plasticity* which turns out to much greater than it used to be believed, which

currently is redefining important brain research paradigms such as brain function narrow specialisation (Schwartz and Begley, 2003).

Before we introduce the mainstream hypothesis by neurologists on how consciousness happens and is implemented in the brain let us make a general introduction that unlike in philosophy of mind where unity of consciousness is underlined, in neurology there is a tendency to look at consciousness as process or even a compound of processes to which more than one area in the brain contributes. Similarly in case of the concept knowledge there is a tendency to look at it as a process: *process of knowing* rather than an object of possession. Furthermore the sole notion of the consciousness is usually considered as a triad: (i) wakefulness or awareness, (ii) mind, and (iii) self, and often deeper decomposition of these concepts takes place, similarly as it happens with brain regions and their corresponding functions.

Damasio (Damasio, 2010) proposes that normal consciousness is only possible if the three main conditions are fulfilled: (i) a subject is awake, (ii) it has a fully operational mind, which is constituted by brain processes responsible for capturing, memorising and handling all sorts of images (representations of intentional contents), both sensory and inward directed, (iii) last but not least, it has a "sense of self as protagonist of experience".

> "Brain constructs consciousness by generating a self process within an awake mind. (...) Wakefulness and mind are indispensable components of consciousness, but the self is the distinctive element."(Damasio, 2010, p. 180)

The sense of self, the protagonist with the thought of itself, comes in three stages as well: (i) *the protoself* that is formed by images mapped from the body inner part, by so called *primordial feelings* fundamental, spontaneous feelings of a living body that originate at the level of brain stem which is an evolutionarily older part of the brain and the meeting point of the cortical parts of the brain and the spinal cord connecting the body with the brain; (ii) *the core self* that emerges as the protoself gets modified by an interaction with objects of external world within the space of mind; (iii) finally *the autobiographical self* is fully formed when the pulses of core self created by manipulating images of past and anticipated future experiences interacting with protoself are put together in a coherent pattern creating a continuous autobiography across time. Importantly, consciousness should not be equalled to wakefulness, nor to mere mind. Mind is a fundamental space in which protoself interacting with images forming the core self and the basis for emerging the autobiographical self, so it is a part, or a sine qua non condition for consciousness, meanwhile wakefulness determines to which degree the self is present in mind. Under the account of Damasio the qualitative feeling of experience is realised at the level of protoself. The primordial feelings that build up the proself have their particular quality, valence, somewhere along the pleasure-to-pain range but also intensity as the pleasant and unpleasant

states may be stronger or weaker. This is a critical stance, as this means that all the contents of mind are in a way reflected in the protoself, inheriting undoubtedly the feeling, qualitative valence and intensity from it.

Somewhat similarly Panksepp (Panksepp, 1998; Panksepp, 2008) talks about three levels of processes that correspond to the three stages in which full human consciousness arises: (i) primary processes that reflect brain/mind basic functions of biological brain; (ii) secondary processes reflecting learned symbolizations and object relations in the world (which sounds like those standing behind intentionality; (iii) finally tertiary processes, which unlike the previous two might be only found in humans, which stand for autobiographical self, i.e. thoughts about thoughts and emotions. This directly corresponds to the three levels of consciousness that Tulving talked about: (i) *anoetic consciousness*, meaning experience without knowing; (ii) *noetic consciousness*, knowledge-based consciousness; and (iii)*autonoetic consciousness* which is build by autobiographical memories (Tulving, 1985). Panksepp believes that affect is present at each of these three levels, primary process emotions, secondary level emotional processes modulated by learned relations, and higher-order derivative processes that may be only studied in human experience. Importantly, we know relatively much about the primary-process affects based on extensive studies with animal models, much less or very little on higher-level process affects typical for human experience (Panksepp and Campbell, 2010).

Ledoux (LeDoux, 2002) contributes to the discussion on neurological underpinnings of consciousness primarily with his considerably sceptical account of the overall importance of conscious experience compared to the non-conscious brain processes. Ledoux indeed supports the claim that consciousness will be fully explained in terms of brain processes one day but at the same time notes that this discovery will not be a major breakthrough as what is evidently far more important for understanding the human nature are the non-conscious brain processes. These cannot be studied by introspection, solely by brain studying experiments. Consequently he distinguishes two parts of self: (i) *implicit self* which accounts for aspects of self that are not immediately available to consciousness either because of their implicit nature or being not accessible at a given moment, and (ii) *explicit self* which is all that is immediately, consciously accessible to the subject of experience.

Concluding the above introduced contemporary neurological theories of consciousness let us assert that affect plays a fundamental role in building consciousness. It is also important to stress that conscious selves subjectively experience their "I" in the way it is proposed by philosophers of mind, i.e. as a unified field of consciousness, rather than the way neurological account suggests, as multiple brain processes. So from the perspective of an experiencing self there is one conscious field. Still the objective nature of experience is likely to be as brains study evidence tells us, so in the deliberative process, in which the self engages, it takes knowledge contents that are amalgamated with primordial feelings and makes decision based on both

intentional contents and affect that are from the subjective perspective at least inseparable elements of a conscious mental state.

It is clearly visible from the above very brief overview of main theories of consciousness that it is very difficult to say what consciousness is without referring to other concepts such as mental states, experience, being, I-am-ness, etc. which themselves are ambiguous. Consciousness and intentionality are so fundamental concepts in the study of mind that they become axiomatic. This exemplifies the difficulty in approaching and discussing consciousness at all, as the efforts targeted at defining consciousness often ends up in a deadlock of recurrence. For this reason it is reasonable to try to define the phenomenon descriptively, i.e. to enumerate the unquestionable or at least plausible characteristics of consciousness which tell us more about the nature of it. Interestingly, with few examples, modern philosophers of mind tend to agree on the majorities of qualities of human consciousness despite being at odds when it comes to its genesis.

To wrap up, consciousness is the playground for conscious experiencing. The intentionality of mind realizes itself in consciousness and fills itself up with intentional contents. Consciousness in this sense seems equal to mind. Mind is simply consciousness in all of its modes. The totality of mind's intentional contents that are directed towards both outside world and within of the mindbody build up the experience. With such an account of consciousness let us have a look into its main properties. These properties are critical for providing a plausible account of knowledge and later methods for their representation in the information systems.

3.2.2 The qualities of conscious experience

The only possible and available mean to understand consciousness and its significance to purposeful behaviour is to realise how descriptively consciousness is like, what properties it has, what is its structure. Different authors suggest different number of defining properties of consciousness, James (James, 1890) identified 5 "chakactees" (characteristics) of conscious thought, Searle counted up 11 (Searle, 2004), but we would like to concentrate on 6 that are always between those to be mentioned first and are highly significant to the remaining part of the argument.

Qualitativeness (i) and *subjectivity* (ii) are the most significant properties of consciousness and may be considered together. The conscious mental states have a subjective, i.e. experienced by some sort of a subject: an agent, an organism, and qualitative feeling in the sense that it always feels like for this subject to be in that state. What is highly important this feeling apart from taking one of endless possible qualitative forms it can be either pleasant or unpleasant to the subject so that it is always unambiguously liked, disliked or ambivalent. Some authors treat the *qualitative valence* as separate property (e.g. (Searle, 2004)) but to us it seems that it may be considered together with qualitativeness in broader terms. The qualitative

character of conscious mental states is of key importance to our later argument for which reason we will discuss it thoroughly in a separate, following section, and we will come back to it later still on many occasions.

Unity (iii) of consciousness is the third fundamental property of consciousness. Consciousness is unified in two distinct manners, firstly it is unified across time, one mental state follows another and keep consistency over the lifetime of a healthy individual, secondly at each point in time the consciousness is felt unified in the sense that different appreciations do not come separately but as a whole, a textbook illustration of it is that we do not feel separately our shirt on the back and the objects in front of our eyes and the sounds of someone speaking as separate instances but rather as one, unified experience. The property of unity can be illustrated with either of the two following metaphors: a metaphor of "stream of consciousness" or a metaphor of "unified field of consciousness". There is an important difference between these two approaches, the former one assumes that the consciousness is a sequence of discrete mental states equivalent to appreciations of sensations or thoughts meanwhile the latter sees consciousness rather as a fluctuating field in which these appreciations are reflected. Both views have their supporters and opponents, the former is supported for instance by James and contemporaneously by Dennett, the latter by Kant, contemporaneously by Searle (Searle, 2004) and Damasio (Damasio, 2010) to name a few. Similarly, there is an ontological consequence of this difference for consciousness, in first case the consciousness and its unity can be considered illusionary meanwhile in the second the ontological and physical existence (in some sort of neural representation) of consciousness is assumed. The unity of consciousness is a subject of itself, notably a separate book by Michael Tye on the topic appeared delving into this particular quality of human experience (Tye, 2003), where many types of unity are identified and discussed: object, subject, spatial, phenomenal, introspective, Gestalt, neurophysiological and higher-order subject unity.

We will assume after many contemporary philosophers that indeed consciousness is unified in a variety of ways, which adds to the significance of the subjective and qualitative accompaniment to knowledge, because it acts causally on behaviour. It is important to underline however that unity, as all the other properties discussed in this section, is a "perceived" characteristic of consciousness which is confirmed by our introspection. The unity at the level of conscious access does not necessarily must mean its unity at "implementation" brain level. On the contrary, recent results of neurological studies of the brain suggest that the consciousness may be realised in the brain via multiple processes, and self is coming in stages: *protoself*, *core self* and *autobiographical self* as suggested by Damasio (Damasio, 2010). However, it is important to highlight that the fact we feel consciousness as unified does matter, even if it may be an illusion as Dennett believes. As we pay much attention to the subjectivity of the conscious experience the unity of consciousness is important to us because it is subjectively perceived as such, so from the point of view of experience

representation and modelling that is critical, and will be addressed in chapter 4 and we will be coming back to this point on many occasions.

The *intentionality* (iv) of consciousness means the unique capacity of mental states to refer to the external world, to be about objects/states/processes that are external to it. This aboutness allows mental states to be "filled in" with content. These contents can be many forms, including sensations, abstract thought, all sorts of *images* in general, which stay with the mind in the intentionalist relation, i.e. mental conscious states are about the external entities and this is how the content of the mental states is build. These contents can be stored in memory in different formats, both in raw images (visionary, auditory, etc.) and organised with some sort of abstract symbolic representations, mostly with natural language. This is primarily why language cannot embrace the entirety of the mental contents, as it is merely a means of organisation of intentional contents of the mind, still it does very well in describing these contents. But regardless the language can capture entirely the intentional contents or not, it definitely cannot capture what is beyond it, i.e. the qualitative, inner subjective feeling of the subject that accompany the creation of intentional contents of the mind. This is pivotal. The intentional contents of mental states is knowledge, the qualitative feeling of having this mental state to the subject complements knowledge giving rise to experience. As some part of consciousness is always directed towards the inner side of the mind and body of the subject and in parallel towards the external world realising its intentional capacity, and at the same time these two are amalgamated in the unified field of consciousness, both knowledge and its qualitative subjective feeling build up the *unified experience*. We assume it is safe to risk a statement that it is not knowledge equals experience as Einstein proposed, rather consciousness is experience and knowledge is its intentional contents.

The fifth quality of consciousness to be discussed is *attention selectivity* (v). Consciousness is selective in the sense that mental contents that appear in the field attract different level of attention from the self. The self tends to engage in only one selected content at a time over a given attention span and the remaining part is passing in the background. It may be the case however than more contents are given higher attention in parallel, which happens for instance in multitasking, however the universal property of consciousness is that it has a center and periphery, so some contents manage to draw the attention of the self others to a lesser extend. Although there can be significant plasticity in this property of consciousness and capability of managing the focus of attention may vary significantly between individuals, as for instance some people are more capable of multitasking and following two or more threads of thought in parallel.

The interesting part is how attention is governed. Damasio's account of the issue is that a hypothetical mechanism which he labelled "somatic markers hypothesis" applies here, which in short says that each content that appears in the consciousness

is emotionally marked or "highlighted" which allows the mind to decide about the importance of the appreciated stimuli and present them to the attention of the self. This hypothesis will be further discussed in the following section as it is important for the arguments to be presented later in chapter 4.

The sixth property: *plasticity* (vi) will be introduced and discussed at the end of the section providing for a conclusion and interlude to the subsequent parts of the chapter. Besides the aforementioned six characteristics there are also other included by philosophers of mind and cognitive scientists which are less relevant to this thesis however are vital for understanding consciousness as a whole, therefore we will briefly discuss them in the following paragraphs.

Searle (Searle, 2004) distinguishes: *situatedness*, *Gestalt selectivity*, and *mood dependence* as separate properties of consciousness, which is debatable. Importantly the existence and reality of this properties is not to be questioned but merely their separation as independent properties. We see that these are already captured by the six properties discussed earlier.

Situatedness (vii) which is embraces the property of consciousness such that conscious self always know the some basic background, contextual facts, such as the time of the year, the whether, the geographic location, the time of the day, etc, we see as redundant with unity and intentionality, as the context that situatedness encapsulates seems nothing but repeatedly invoked intentional contents about the environment the agent sits in and the continuity of these contents due to the unity of consciousness. Similarly Gestalt selectivity seems a combination of unity, indeed Tye (Tye, 2003) refers to it as Gestalt unity, and attention selectivity.

Gestalt selectivity (viii) is a highly interesting property of conscious mind, presumably also the unconscious mind works in this way but it is not evident, extensively studied by Gestalt psychologists (Köhler, 1929) which can be described as an organisational capacity of the mind aiming at, in plain terms, making sense of the intentional contents it captures. This sense making is usually about building up a "bigger" and complete concepts from the captured sensory data, fitting the newly captured sensory data into these bigger concepts, which is a mechanism that serves the organisation of sensory data and mental intentional contents into meaningful concepts and more elaborate mental stories persistent overtime. Gestalt psychologists used behavioural tests to prove this in which they used purposive crafted pictures that represented at the same time two different images, of which the classic textbook example is a black and white picture that can appear both as two faces directed towards each other or a vase. Depending on which of the two images the mind concentrates on the one is captured by the conscious mind. The interesting properties of the conscious mind is that it can swap between the two but it is hard or impossible to "see" or, better said, experience both at the same time. The above description of the phenomena shall serve as sufficient justification why we reckon these properties as a derivative of broadly understood unity of consciousness and selectivity of attention.

Finally, *mood dependence* (ix) fits under the scope of qualitativeness and will be discussed together with it in the following section.

Another proposition by Searle is to tell apart the *active/passive* (x) consciousness as states of consciousness activated during engagement in a voluntary action and passive perception respectively. Although, as he rightly stresses, there is a clear difference between voluntarily raising an arm as part of a conscious act, and having an arm raised by someone who stimulates relevant part of motor cortex, an example taken from the famous experiments by neurosurgeon Wilder Penfield (Penfield, 1975) [2], we would rather think of a voluntary action mechanism not as an integral part of consciousness but rather as a mechanism to which consciousness is a "background environment". Consciousness is necessary for voluntary purposeful behaviour and is the background for it but it is not consciousness itself.

Mode dynamism (xi) is a property that has a more indirect consequence to our considerations and is about ability of consciousness to be in an "on" and "off" mode and some other modes "in between". The modes in question include: normal wakefulness, dreaming, deep sleep without dreams, and different forms of coma. As Damasio rightly says (Damasio, 2010) the boundaries between these fairly stable and well studied states of consciousness is not so sharp. We can experience this ourselves when our consciousness shifts from one to another. Being tired, sleepy, falling asleep, awaking, loosing consciousness are all intermediary states in which our state of consciousness is largely undefined, which we experience and define as passing out or coming to. Studying these processes is one of the way towards understanding consciousness and its building blocks. Damasio hypothesises that the way consciousness is put together and decomposed each time consciousness changes its operating mode suggests that consciousness relies on more than one core location or system in the brain and it is rather a process than an object. He proposes that consciousness is composed of a few building blocks which get activated non-simultaneously, provoking the particular, awkward feeling to it, lasting until all are finally rightly "lit-up". But more importantly, it may happen that only some of these blocks get activated while others remain inactive for a longer period of time as consequence of drugs or physical interference, a knock in the head or solar plexus for instance, which provides some clues on how consciousness is engineered in the brain.

These findings are important for us only indirectly, i.e. the property itself does not impact much our later considerations and the frameworks elaborated in chapter 4, as we have already rationed the scope of our focus to the normal, awake consciousness state, but it is interesting as far as it provides a way towards understanding the nature

[2]Penfield, who was searching for causes of epileptic seizures and regions in the brain involved in the disorder, performed experiments that involved stimulating with electrodes the surface of exposed patient's brain trying to thus identify the regions that triggered the seizures. When he stimulated the motor cortex he could provoke involuntary movements of patient's limbs. However the patients described the corresponding conscious experience provoked by this stimulation with these words: "I didn't do that, you did it!"

of consciousness in general. And so, Damasio's conclusions form his investigations and theoretical considerations are that we can look on consciousness from two distinct vantage points which results in two types of self: self as observer and, younger from the evolutionary perspective, self as knower. The former appreciates a dynamic object constituted by workings of minds, traits of behaviour, and history of life, the latter is the process that gives a focus to our experiences and lets us reflect on those experiences (Damasio, 2010, p. 8). The consequences of this distinction will be discussed in the following section when we look into the role of emotion in creating self.

The emergence of self from consciousness is usually taken as one of the characteristics of consciousness which we shall refer to as *Iamness* (xii). Nagel (Nagel, 1974) in the essay "What it is like to be a bat?" noticed that conscious experience defines act of being, underlying thus the very subjective character of experience. Mind is that which thinks and experiences, thus thinking and experiencing defines a minded creature, including a human being (Graham, 1998). Nagel was one of the first contemporary IA thinkers to emphasize the fundamental significance of conscious experience to constituting being. One can say that something is only when there is something that feels like being that thing. Therefore in order to be able to say "I am" there must be consciousness. This immanent Iamness of consciousness is where ontology meets epistemology. If we suppose for a while that consciousness in a much more broad sense is pure intentionality and assume that beings which do not have enough elaborate brains to have consciousness of a sort that has been discussed throughout this section, still can have some more primitive capabilities of intentionality, which are abundant in even simplest single cell organisms, but also in could be found in the domain of plethora where lifeless objects are bound with laws of physics, we indeed can risk stating that existence without intentionality, i.e. ability of one entity to be about another, would not make any sense nor make any difference therefore had no *meaning*. Although the Iamness of consciousness unquestionably touches the most fundamental philosophical questions, for the purposes of this thesis it is sufficient that we consider it as element of subjectivity and unity of experience.

Finally, we would like to add to the list of considered properties of consciousness the *plasticity* (xiii). The plasticity of the brain is one of the topical issues in current neurocognitive and brain studies (Kolb, 1995; Shaw, McEachern, and McEachern, 2001; Schwartz and Begley, 2002). As brain manifests incredible flexibility and adaptability in many of its functional areas very likely this capability is also visible in consciousness, to which some evidence of introspective nature can be provided. The good example confirming the plasticity of consciousness are those that demonstrate that certain otherwise fixed, or common for many individuals, properties of it can be altered by practice or experience over time. Such an example is the human capacity to multi-task, i.e. concentrating on doing more than one thing at a time. Capability of multitasking questions the selectiveness of consciousness, and can be gained by

practice. Especially nowadays and in younger generation, listening to music and working/learning is a typical practice. The number of life management supporting devices such as mobile electronic assistants, phones, tablets, portable computers that allow us to be "always on" impose on its users a requirement of multitasking. Heavy users of this type of life-assisting machinery are forced to concentrate attention on more than one thing and with time develop a capability to maintain a few streams or consciousness or rather more than one threat within it. Another example is training concentration and voluntarily control consciousness in meditation. In Ken Wilber astonished the scientific community presenting tapes with recordings of his encephalographic images of his brain activity while he meditated, showing that he can voluntarily control mode of his consciousness reflected on EEG results with alteration of electrical activity of his brain. What Wilber demonstrated was that he can voluntarily control which of basic waveforms (the alpha, beta, theta, and delta) his brain emits, ability that is beyond capacity of a normal, healthy brain, which always emits waveforms that are typical of a given brain state or a stage of development, e.g. delta are emitted normally by young children or during deep sleep, beta after taking certain drugs, alpha are present only when one is awake but with eyes closed, theta are common for children and young adults (Wilber, 1977b; Cahn and Polich, 2006; Shapiro Jr and Walsh, 2008; Wilber, 1977a).

This is just one of may examples how mindful meditation can result in brain/consciousness plasticity. Other are Schwartz's (Schwartz and Begley, 2002) research at UCLA in obsessive compulsive disorder treatment by means of mindful meditation, or Kabat-Zinn's work on treating stress disorders with the same technique (Kabat-Zinn, 1990; Davidson, Scherer, and Goldsmith, 2003). Another comes from neuro-plasticity pioneer and innovator Dr Michael Merzenich who worked out and patented many training methodologies that leverage brain plasticity to assist people with cognitive dysfunctions, in particular children's language, learning and reading problems (Buonomano and Merzenich, 1998; Merzenich et al., 1996).

We deeply believe that these capacities of human brain are yet to be fully explored, and its limits remain largely a mystery for us. Likely achieving certain states of consciousness and the brain which are now possible solely by alterations with drugs or surgery can be achieved by practice directed inwards, introspectively. This however is a topic that falls beyond the scope of the thesis and served solely to provide evidence for consciousness plasticity.

From the point of view of our later arguments in chapter 4 the plasticity of the brain, consciousness in particular, is somewhat problematic. The plasticity means that the mind/brain is a highly adaptive therefore dynamic machinery, for which reason it is hard to be captured and expressed in hard formal terms. Whatever a description of this machinery, or its parts, in whatever formal system must include structural plasticity itself. Learning, machine learning may be a solution here though apparently the nature of learning this plasticity mandates is not necessary the

learning mechanism captured by learning algorithms. We will come back to this discussion in Chapter 4, however let us highlight already here that the yet largely unexplained plasticity of the brain, consequently consciousness as well, poses one of the major obstacles and limits on the representation of experience in artificial systems that by their nature lack this plasticity. On the other hand studying further brain plasticity may provide for a better understanding of the indeterministic nature of brain and consciousness manifested in free action. We will come back to this point later speculating in the context of a discussion on the practical reason that brain plasticity may serve as the entry point for explaining the phenomenon of free will.

In the next subsection we will concentrate on the qualitative dimension of experience. We will look at some most important question: What subjectively colours consciousness, and consequently what colours conscious experience to be able to take an account of how does it impact knowledge and its application in rational behaviour.

3.2.3 Temporal dimension of conscious experience

In this section we will introduce an important quality of conscious experience that results directly from the characteristics of biological brainmind functions: its volatility in time.

The capacity to memorize mental/brain states and make references across time between those states is one of the most, if not the most important capacity of human brain which makes us unique from other mammals and primates. Many thought leaders in the study of memory, such as Eric Kandel (Kandel, 2006) and Daniel L. Schacter (Schacter, 2002) do believe that extraordinary memory capacities of human brain makes us who we are and let us lead in the race of evolution. These capacities decide about our intelligence and adaptation abilities, as impediments in memory building caused by various types of diseases lead to serious limitations in basic mental capabilities such as creativity, decision-making, learning and cognition. As Schacter rightly puts it "Memory is so fundamental to the operation of the brain and mind that students of the topic could be forgiven for feeling that their object of study is perhaps the most central in all of cognitive neuroscience" (Gazzaniga, 2009, p. 655).

For the above reason we insisted on introducing this aspect into the definition of experience, emphasizing that experience is build of *remembered* states of mind. Regardless how spectacular the ability of a biological mind to remember is, we must not forget that memory machinery is not flawless. Due to economic and practical reasons imposed by evolution not each and every mental state can be stored in the memory for lifetime. Our memories are incomplete, fragile, sometimes corrupt (false), may be temporarily unavailable, can be permanently be kept away from the access of consciousness (implicit memories). This has a direct influence on experience.

Experience directly inherits these capacities from memory, in fact we can risk a statement that memory as phenomenon, not so much as capacity, equals to experience. For this reason there can be no sound account of experience that does not address processes of remembering and forgetting, which determine the temporal dimension of experience. These processes are studied by cognitive neuroscientists investigating memory processes, which usually include: (i) encoding, (ii)storage and (iii)consolidation, and (iv)retrieval. Furthermore, from the cognitive functional perspective a couple of memory forms can be considered: working, episodic, semantic, priming and procedural memory, each governed by different functional nuclei or rather interlinked neuronal circuits, which shows that memory is not a single phenomenon or entity but rather a complex compound of brain sub-systems.

Importantly however, as we have made a point in previous sections when we discussed consciousness from the perspective of the philosophy of mind, an important quality of conscious experience is *unity*. It must be underlined that memory processes are in vast part unconscious processes, and only the end result of these processing is accessible to us consciously in thinking and deliberation. We can distinguish therefore between implicit and explicit memories. In fact, a vast part of learning, i.e. encoding, storing and consolidating memories, is carried out by the brain in the unconscious. Our interest here is primarily focused on explicit memories that can be voluntarily retrieved in consciousness.

The objective of this subsection is not to provide a detailed account of how memory functions however, instead we want to emphasize two highly important points:

1. Experience being totally dependant on memory over time is subject to all the four major memory processes, consequently it is prone to forgetting which is a fact widely neglected in mainstream KR methods, therefore there is a constant need for definition of algorithms that could emulate fundamental memory processes in artificial systems;

2. There is reach scientific evidence to that memory is strongly dependant on emotion, and that amygdala, a group of nuclei playing key role in emotional management in the brain, modulates memory processes, it is therefore critical for viable algorithms emulating memory processes in information systems to take affective accompaniment to knowledge into account, which in turn requires solid methods for affect representation in these systems.

The remaining part of the section will provide a brief overview of characteristics of memory process that shapes most significantly the temporal dimension of experience, notably: (i) the role of past memories in shaping the present and the future (ii) the impact of emotions and feelings on memory.

Bowker inspired by Lyell, who was one of the classics to claim that history does not exist as it just constitutes part of the present (Lyell, 1837), aptly gives us clues

on how important it is to look on the past experiences as part of the presence by noting:

> "Our reading of the past is generically under the description of the present set of entities and phenomena (...). Past time is the same as present time; past entities of the same order as present entities." (Bowker, 2005, p. 227).

This assertion is strongly supported by neurocognitive account of brainmind, memory in particular. It is not only that we use our current cognitive capacities to deliberate about the past experiences, but also the past serves us as a founding base for anticipating and simulating future events. First of all it must be noted that in general the memory works in a way that the present mind states (brain maps build constantly in real time) are encoded into a form of hash codes that serve in the recall process to recreate the remembered map in the brain. Furthermore the memories about objects are rather the record of multiple consequences of the organisms interactions with them than the pictorial representation of structural details, for which reason the memories about objects normally involve sensorimotor patterns associated with the seeing, touching, manipulating the object as well as patterns related to the triggering of emotions and feelings relative to the object and previously acquired memories about the object. The above is suggested by the Convergence-Divergence Zones (CDZ) theory of memory proposed by Damasio (Damasio, 2010), which accounts that brain does not record the complete representations of the overall mappings of neuron activities in different parts of the brain in the given instance, but rather the coincidence of signals from neurones linked to the given map. When the memories are recalled a mechanism of *time-locked retro-activation* is started in which components of the map are retro-activated roughly simultaneously. This results in an intrinsic capacity of memory machinery to store information not so about fragmentary, isolated elements of reality but rather entire momentary instances of unified field of consciousness, and the entire context of interaction with an object (event) is both remembered and recalled. However, the recall as it is made based on the coincidental "hashed" neuronal maps instead of complete representations of maps, may not be precise and is prone to error. This explains the contextual unity of consciousness which persist over time and the imprecision of memory (forgetting process) which starts already at the encoding stage and is dubbed by the instability of memories over time.

The CDZ theory is an exemplification of a contemporary neurocognitive approach to memorizing which suggest that remembering is not like reading out a stored file, but rather it is a *reconstructive* process in which past and current experience mingle with new information. (Gazzaniga, 2009, p. 691). The research on memory re-consolidation, vastly advanced over recent decade, has provided us with important insights into the process of forgetting. Traditionally it has been believed that long-term memories are built in steps, first they exist in a labile (unstable) state and

afterwords they get fixed in the process called consolidation. It has been shown recently that each time during recall memories enter into liable state even those that have been consolidated already.

One of the many studies that support this account was conducted by Ledoux and his team, on animal models: rats. Ledoux investigating the classical conditioned learning in rats discovered that the moment a conditioned memory is recalled by re-exposure to the conditioning stimulus the memory system gets destabilized for several hours and if during this period the memory storage process is altered either behaviourally or with drugs the memory does not go back in properly and it is lost.

Another important recent discovery in memory research is the one made by Schacter and colleagues (Schacter and Addis, 2007) about the role of constructive memory in the simulation of future events. A series of converging results of recently reported studies suggest that the brain uses the same neuronal machinery to imagining future possible events that is normally applied to remembering past events. This nicely fits in the convergence-divergence approach and suggests that deliberation about future events, normally also present in decision making depends tightly on past experience, furthermore adds arguments to the claim that creativity is about reconfiguration of existing facts which leads to new emergent qualities that stand behind the novelty.

All the above suggests that biological memory is very much unlike the computer functionalism wished to imply, i.e. that memories are stored in the brain like files in a computer system and reproduced during recall. The old paradigms of knowledge representation deeply rooted in computer functionalist tradition fail to embrace this very basic characteristic of biological machinery of minded organisms for knowledge acquiring and processing. There seem to exist no past experiences only their reconstructions in the present. So the past matters only to the extent it reappears in the present experiences. Furthermore the past, the present and the future use the same neuronal processing machinery and apparently are mingled in a conscious field of consciousness according to a pattern that is not fully understood yet. However as experience depends on memory it is evident that it is subject to forgetting and distortion, which should be taken into account when representation of experience is to be done in information systems.

Apart from the overall importance of memory processes to experience, there is a special dependence between memory and emotion, which suggests that affective dimension of experience matters not only to the past and present but also future, consequently has profound impact on planing future actions, thus conscious behaviour.

As a general principle, emotions amplify memories, as it has been proven by investigations led by McGaugh (McGaugh, 2000; McGaugh, 2004) that amygdala influences consolidation of emotionally arousing experiences. The amygdala activity during experiencing emotionally arousing stimuli, regardless positive or negative, proves to impact positively the consolidation of memory, so that the degree of amyg-

dala activation during encoding correlates with long-term memory. The emotional arousal also helps in better recalling of encoded memories and augments their subjective vividness. One of the way it is achieved is that amygdala modulating the emotional arousal, being well connected to sensory cortices amplifies the activity of this cortices which leads to better detection and intensification of attention.

Furthermore, orbitofrontal cortex involved in some types of emotions modulates conceptual processes. Emotionally marked memories are also more likely to benefit from the consolidation amplifying effect of sleep on memory (Gazzaniga, 2009, p. 730). It is also known that memorized states of mind are more easily retrieved form memory when there is a match between the emotional state at the time of memorizing and at the time of reconstruction, so that one is more likely to remember negative experiences when is in a bad mood and positive during feeling of pleasure or other positive excitation (Bower, 1981).

There are important exceptions however to these principal general rules. It has been found out that stress impairs explicit memory by affecting the functioning of hippocampus (Sapolsky, 1996). [3]

In conclusion, there is strong evidence that memory, which provides for the long-term unity of experience and its intentional contents (knowledge) depends on the subjective, affective quality of experience. This provides for yet another argument supporting our claim that representation of affective dimension of mental phenomena is indispensable for satisfactory emulation of knowledge, experience and behaviour of living organisms in artificial systems.

3.2.4 The subjective component of conscious experience

One of the most difficult and puzzling problems about consciousness, is its subjective quality. This problem is plainly and very aptly formulated by Richard Dawkins when he was asked by Charlie Rose the question that Rose always asks the interviewed scientists, i.e. "What is the one question that you most want to see answered?", to which Dawkins replies:

> "How does subjective consciousness work... how does it evolve... what is going on when I have my own private feelings and you have your own private feelings? What happens when I see something red... what is it that makes the redness, what is it that makes the smell of onions? What is it that gives the subjective sensation that I know I have, and I suspect you have, but I can never know what is going on inside your head" (Dawkins, 2005) [4]

[3] Stress has also detrimental effect on sound decision making as it affects the prefrontal cortex.

[4] Based on author's own transcript from the online podcast at charlierose.com, '...' indicate pauses.

Dawkins captures all three elements that make this question one of *the* questions that contemporary science confronts: (i) if and how subjective experience can be quantified, (ii) if and how it can be described in objective, tangible terms, what is the corresponding brain processes behind these subjective states, (iii) how does it evolve, (iv) can it be put in inter-subjective context, in other words can it be compared between individuals? These are central questions to this thesis as well because if they only result in negative answers this objective could never be attained. Although indeed the above formulated problem is still challenging scientists across disciplines it would be definitely unfair to state that no progress has been made towards disentangling the puzzle of experience subjectivity. We risk to state that the progress has been suffcient to come up with preliminary, yet satisfactory, means of representation and intra-subjective comparison of subjective states of consciousness, although in general we are far from in-depth understanding of the phenomenon. In philosophical and AI literature this subject has been extensively explored in the debate over *qualia*. The debate primarily focued on the quantifiability, universality and inter-subjectivity of subjective experiences. This topic is important to philosophy of mind because subjectivity of consciousness is one of the key arguements for the opponents of the phisicalist account of the mind-body problem. The philosophical debate on qualia tackles exactly the essence of the subject matter: the qualitative conscious experience, for which reason this debate shall be briefed and used for review of arguments that may be brought up by phisicalists who perceive qualia as an unrelevant concept and thus partly at least object the importance of the qualitative dimension of experiernce reducing it to purely phisical properties. Qualia could be seen as experience quanta, and are defined as experiential properties of sensations, feelings, perceptions, thoughts, desires, etc., in other words qualia include what is like to have experiential mental states. Taking the most commonly used example, also featured in the earlier quote by Dawkins, a quale represents what is it like to see red for instance, in other words what makes redness as perceived by experiencing bodymind. This represents the most general and intuitive understanding of qualia but as there are many definitions of the term different of them emphasize different constituitive features of qualia, among which the most important are (i) subjective phenomenal character; (ii) universalism, but qualia are universals in the sense that they can be recognized from one experience to another experience and are not the properties of any object (Lewis, 1929); (iii) qualitative feeling, which constitutes the very essence of a quale: the associated quality of experience (Chalmers, 1996) [5]; (iv) ineffable, as they cannot be communicated, and can be fully appreciated only via direct experience, at least according to Dennett (Dennett, 1988; Dennett, 1991); (v) intrinsic, that is they are non-intentional, non-relational, so that they remain the same regardless experience relates to different objects, (vi) introspectively

[5]for this reason qualia are also referred to as 'qualitative feels' or 'phenomenal qualities'

accessible in consciousness, so that it is possible to know that one experiences a quale, which to some theorists including Dennett and visibly Dawkins [6] by consequence means that they are (vii) private, in the sense that it is impossible to compare qualia interpersonally and also (viii) incorrigible, so that the subject cannot be mistaken about his qualia; lastly (ix) non-physical, that is they do not have any correlate in the physical world, yet (x) irreducible and (xi) non-physical, that is they cannot be either reduced to some lower level processess or phisical properties. Qualia are usually thought of being properties of certain states of mind, not all. Among those that are most commonly mentioned are perceptions and bodily sensations, we would suggest to stict to all intentional mind states directed outwards the intentional mind. Strawson (Strawson, 1994) proposes that raw thoughts themselves can also have qualia, as there is a subjective feeling to understanding a sentence or conducting some thought argument. As Tye rightly notices such a position is debatable as the phenomenal aspects of understanding derive rather from linguistic or verbal images. (Tye, 2009) Similarily desires and emotions do not have qualia as such, they rather build up qualia for perceived objects. However, once the self realizes that it is under some emotion this sole fact becomes a belief and forms intentional content. We will elaborate on this later on. It is common that proponents and opponents of the existance of qualia adopt different definitions of qualia depending on that serves better their argumental purposes. In particular qualia opponents tend to define the term more narrowly, anchoring their attack possition on qualia in the features from the second half of the above list, namely: ineffable, intrinsic, private, non-physical and irreducible character of qualia as well as their universalism, which does not mean that they question the existance of subjective, qualitative feelings of experience, which qualia in a more broad sense are in fact (Tye, 2009). This remark is trully important for us, as the view that there are no subjective feelings to mental states is marginal, would require to belive that *zobies*, creatures that are like us but lack subjective, phenomenal experience, could, at least hypothetically exist. The dispute in fact concentrates on how this subjective dimension of experience is, and how much it can be said about these qualities of experience in first place, rather than on whether this dimension of experience is real, as some of the readings of the texts by phisicalists philosophers opposing the qualia concept, such as Dennett, may suggest. In conclusions to this subsection we will provide our account of qualia and comment on their properties that are not necessary and on which qualia opponents are missing the point. Below, to present some of the main arguments of both sides of the debate, we will briefly go through the main line of criticism as presented by one of the most recognised philosophers of mind and proponents of reductionism Daniel Dennett as formulated in (Dennett, 1988; Dennett, 1991). Dennett in his widely cited article "Quining qualia" (Dennett, 1988) questions not so the very existence of subjective

[6] See the openning quote of this subsection.

experience, but rather the relevance and importance of the concept of qualia and their alleged *special* properties, asserting that:

> "conscious experience has no properties that are special in any of the ways qualia have been supposed to be special"

. One in all, his argument leads to the conclusion that qualia defined as ineffable, intrinsic, private and directly apprehensible properties of experience, do not exist. Instead, there are:

> "relatively or practically ineffable pubic properties we can refer to indirectly via reference to our private property detectors - private only in the sense of idiosyncratic."

He notices that the proponents of this special properties of qualia revert to intuition while building their arguments so he decides to use so called *intuition pumps* that are properly crafted thought experiments, which would help him undermine the intuitive 'psedo-argumentation' used by qualia believers. This is very symptomatic about the debate on qualia, which is based on considering different sorts of thought experiments or other conceptual structures without reference to empirical data rather referring to introspection and common sense. This makes the conclusions vulnerable to the allegation that these thought experiments are nothing but philosophical puzzles that do not bring new knowledge, however this is such the methodology in philosophy.

The first of the fifteen intuition pumps put forward by Dennett is (i) *watching you eat cauliflower* which undermines the validity of the assumption that the quale: "how something tastes to X at time t" can be separated from the complete context of the experiencing what X is doing or thinking. Although it seems plausible to say that A tastes to X differently at time t_1 than at time t_2 or comparing these two qualia with that corresponding to the taste of A to Y at the time t_1 it is wrong as it presuposes that we can consider these quales in separation to all the rest which is going on (Dennett, 1988, p. 384). Although conceptually it is sound we may respond that as the past experience is remembered the introspective comparison could be possible. One of the central problems with qualia is that they appear not comparable in inter-subjective terms. This results in that it is not possible neither to validate nor falsify the claim that blue colour appears to everybody in the same way, as we learn what blue color is by perceiving the same blue objects. With the term quale at hand it could be rephrased that we cannot tell if the quale of experiencing the blue colour is the same for all people. To discuss this feature of qualia a classic thought experiments called (ii)*inverted spectrum* is used(Locke, 1801). In the classical form it calls that we imagine that one day we wake up and realise that all colours has been inverted, although we cannot discover any changes to the physical qualities of the objects we perceive nor to our brain. The lesson from this story for a qualia proponents is that if we can imagine such a spectrum inversion possible qualia must

exist and be non-physical. The original thought experiment by Locke has gained more sophisticated variations as the debate on qualia has continued, arriving at a point to the version proposed by Block *block1990* which goes as follows. Let us imagine a planet on which all colours are inverted, so that sky is yellow and lemons are blue, still the inhabitants name the colours of all the objects the same way as we normally do, so the sky is still blue in their natural language. If one is anaesthesia and undergoes a surgery during which "color inverting lenses" are put in one's eyes after which one is placed in the Inverted Earth no difference can be noticed. So there does not have to be a direct link between the object of experience and qualia. Dennett in turn rejects not only inter-subjective comparability but also intra-subjective comparability of qualia. For this he uses the intuition pump *the Brainstorm machine* which supposes that there is a wonderful machine which allows person A to experience the visual experiences of a person B, but when the machine is on it turns out that A sees all colours inverted and is supposed to realize that B has different qualia. But, what would happen if the technician turns the plug by 180° and reverts A 's to its "normal" state of experience? How would it be possible to judge which of the positions of the plug is correct? This could not be possible, in other words we could not calibrate such a machine and we could not know if qualia of A and B are the same. (Dennett, 1988, p. 387). As for intra-subjective comparison of qualia, Dennett provides a counterargument to the *intra-personal inverted spectrum* which proposes that if one waked up in the morning and discovered that the colours are inverted, meanwhile nobody else notices any difference at all this would have made you think that you must have had an inversion of colour qualia (Lycan, 1973). Dennett's counterargument is an intuition pump he calls *alternative neurosurgery*, which consists in an observation that this colour qualia swap could be achieved with two kinds of operations: (i) by changing the qualia themselves, so the way colours appear to us, or (ii) by altering all our past experiences related to colours. Again, one could not tell which of the either two operation was performed. Consequently Dennett concludes:

> "If there are qualia, they are even less accessible to our ken than we had thought. Not only are the classical inter-subjective comparisons impossible (as the Brainstorm machine shows), but we cannot tell in our own cases whether our qualia have been inverted, at least not by introspection." (Dennett, 1988, p. 389)

This conclusion brings Dennett to assertion that if we cannot compare qualia neither interpersonally nor at interpersonal level it does not make sense to talk about qualia at all. The attack continues by stating that in fact we can not say anything objective about them, they are intrinsic and thus atomic and unanalysable, and consequently inexpressible in language, and as there is nothing that can be asserted about the subjectivity representing qualia they are just pseudo-concept, to put it in

Wittgenstanian terms, and Dennett concludes that qualia are not even something about which nothing can be said, they are a philosophical pseudo-term that refers to no properties or features at all. (Dennett, 1988, p. 387). Dennett uses here the following argument by Wittgenstein:

> "The thing in the box has no place in the language-game at all; not even as a something; for the box might even be empty. - No, one can 'divide through' by the thing in the box; it cancels out, whatever it is (p.100) („) It is not a something, but not a nothing either! The conclusion was only that a nothing would serve just as well as a something about which noting could be said." (Wittgenstein, 1958, p. 91-100)

This argument appears weaker if one ceases to regard language as the ultimate carrier of intentional contents of mind. Given we accept the existence of non-linguistic forms of intentionality we provide space for that something can be expressed with this forms about subjective dimension of experience. Still the way of interpretation holds also for another thought experiment that we propose *think about tomato* which is somewhat related to the earlier discussed *watching you eat cauliflower*. The intuition pump is about imagining what eating a tomato is like. If qualia existed in a universal, separate form, abstracted from objects of experience, as products of philosophical distillation, as puts it Dennett, we could not do it without recalling specific contexts from our memory in which we were eating a tomato earlier in our life. The specific events associated with tomatoes may be various, such as our last visit to the marketplace, tomatoes eaten up in our soup last night, the legendary tomato juice our grandmother used to prepare, etc. In this intro-spectral thought exercise one can quickly realize that tomato as an abstract entity is entirely qualitatively empty, it has no subjective, qualitative value, means nothing to us. It is only due to the recalled memories of specific events or thoughts associated with tomates by our past experience that allows us to build up in our mind a state which could be like eating a tomato. In this respect Dennett is right stating that qualia can not exist in isolation from the full context of experience. The direction of thinking that takes Dennett is as follows: if qualia are in fact derivatives of experience one should analyse the experience, not qualia, which cannot be even properly defined, there are no qualia there are only states of consciousness that emerge in interaction with the external world.

Every contemporary philosopher of mind provides his own account of qualia and heterogeneity of views is unsurprisingly high as this is an important element of the discussion on the body-mind problem. There is not enough space to provide a decent overview of all the accounts but Dennett was taken as an example of one of the principal opponents to the concept. In this debate we take the side of moderate representationalists (*Ten problems of consciousness: a representational theory of the phenomenal mind*; Tye, 2000) that accounts for qualia as qualities of experiences represented in the mind, so that qualia are representations in consciousness of objects

in the world not phenomenal objects themselves. This view is shared by a wide group of contemporary philosophers including Tye, Chalmers (Chalmers, 2004), Dretske (Dretske, 1995), Searle (Searle, 2004), Siewert (Siewert, 1998) in spite of differences in fine-grained theoretical details.

Representationalism accounts for qualia as equal to certain representational *contents* or as certain representational *properties* of experiences, which we favour. Siewert and Searle find this properties as irreducible. Tye draws an viable analogy between the relationship of qualia and experience on the one hand and meaning of a word and the word itself on the other. Indeed, the view on qualia as *the meaning of experience* greatly appeals to us, however Tye and Byrne make one step further (Byrne and Tye, 2006) asserting that "qualia (like meanings) ain't in the head" so they are relational contents fixed at least partly by external relations between agent and its environment. This assertion we find debatable, we rather consider qualia as subjective phenomena that can be inter-subjectively related and indeed depends on the external entities being represented in the consciousness but their location is in the agent's head. Similarly Searle as proponent of the consciousness unity discussed in previous sections believes that as consciousness functions as a continuous unified process it is not possible to account for qualia as atomistic, quantifiable phenomena but rather as a non discrete compound of qualitative value of experience at a given point in time. However Searle supports the claim that each state of consciousness carries something that is like to have this conscious state, which was discussed earlier in detail. Consequently Searle remains strong supporter of qualia but not as universals but as qualities of conscious experience (Searle, 1999). It is not important to us if qualia have objective existence of any form. It is pivotal for us that they cannot be questioned at the subjective level, which we know based on our individual, private introspection. The important take-away from the earlier discussion for us is that for each, even tiniest experiential state that can be recognised in the consciousness, a qualitative value can be assigned by the experiencing subject, the experiencing mindbody. We also believe, as many contemporary philosophers of mind, that these states can be inter-subjectively compared but we currently lack relevant concepts and ways for doing so, which is an intuitive assertion which at present cannot be objectively verified, irrespective of Dennett's strong opposition to such settlement. He strongly states:

> "Indeed, a subject's 'introspective' convictions will generally be worse evidence than what outside observers can gather. For if our suject is - as most are - a 'naive subject', unacquainted with statistical data about his own case or similar cases, his immediate, frank judgements are, evidentially, like any naive observer's perceptual judgements about factors in the outside world."

This argument is strong and valid, however we cannot afford in the discussion we are conducting here for strict attachment to pure methodology if we want to make any progress in understanding and also exploiting the subjective component of experience. It is alos important to note that this argument cease to be valid if we consider only the subjective, not inter-subjective realm of experience. If we want to understand the phenomena such as human behaviour, intenlligence and motivation we cannot deprive ourselves of the concepts that are from within the world of subjectivity. Qualia are not ontological beings such that there is a universal quale that says what is the feeling of redness universally. However qualia, more precisely the subjective feelings of consciousness, are real phenomena, they exist in subjective consciousness and they matter as they impact agent's behaviour. The closest account of qualia to that we accept is that proposed by moderate representationalist so we rest on the following properties of qualia: (i) are accessible to introspection, (ii) can be different for the same representational contents of the experience but only at different points in time or different contexts, so one may have different qualia for the same object at different points in time or in different contexts, (iii) are mental correlate to some physical properties of objects in the world, (iv) they build up the subjective dimension of experience, (v) they depend on memory, are memorized together with the correlated object and context. Some of the problems with qualia could be solved by taking a dynamic perspective on the phenomenon, as provided above in the 'tomato intuition pump'. Qualia are therefore in our understanding nothing but subjective dimension of experience accumulated, i.e. remembered, over time associated to given stimuli. It is subject to learning, for instance classical conditioning, so qualia are dynamic phenomena that are shaped in the process of repeated concurrence of a stimuli, accompanying stimuli and compound feelings these stimuli provoke. This feelings are the products of constant brain mapping of the body state. So qualia are a sort of representations of external world in the internal world of a minded agent. Qualia in that sense are not inter-subjective phenomena, they exists in the sense that conscious experience has a qualitative value determined by affect, primordial feelings and feelings of emotions to be more precise, available only to the experiencing self. Qualia in this sense are plastic, they can change overtime along with experience. The quale of experiencing redness is not a universal quality attached to the redness, rather it is the qualitative value of experiencing redness by a particular agent. In this sense we could use the term qualia for encapsulating the qualitative, affective component of experiential, conscious mental states, affective quanta of experience. Importantly however if a way for measuring and representing feelings is found out, for instance based on empirical observation of their phisical representations in the brain for instance, then qualia will turn out to be intersubjectively comparable. Contrary to the last assertion there exist a problem of the so called "explanatory gap" (Levine, 2001), which is that as we have access to qualia only via conscious introspection it will ever be difficult to match it with any objective processes that

may be going on in the brain that correspond exactly to the considered quale. Even if we come up with powerful technology with which we would be able to study every little detail of how the CNS functions we will never be able to brige the gap between this apparently different two realms of inner mental states and objective physical reality. This presents to us the the classic problem of dualism. The clearcut solutions comes form phylosophers who reject dualism, like Searle who without questioning the gap believes that some physical qualities or states simply have irreducibly subjective nature (Searle, 2004). Other possible approach is to accept that the gap will never be bridged (Chalmers, 1996) or hope for that some day it will be bridged when proper concepts we lack today are worked out (Nagel, 1974). Yet another which we want to pursue is that the gap is real but it is not so important, so we should not make it wider and more significant that it appears from the pragmatic perspective. This is close to the approach of Searle, though he emphasizes the gap pretty strongly, yet notices that the gap is a consequence of misguiding concepts inherited from Cartesian dualism. As acknowledging that the gap exist has no influence on how things are in the world nor prompts any modification for theories of consciousness the gap does not fill the bill.

3.3 Affective quality of conscious experience

The objective of this section is to provide evidence to the claim that emotions are immanent component of conscious experience and complement the cognitive dimension of experience with the qualitative subjective value. To this end we will take a closer look at the role that emotions and feelings play in shaping consciousness, how they fit in the overall picture of experiencing self. Later we will characterise emotions and feelings to identify impact points on consciousness and purposeful behaviour the same way we have done for consciousness in general. Finally we will take a look on the capacity of language to express affective dimension of experience, which will be a prelude to the final part of this chapter in which we will try to provide an account of agent rationality that includes the emotional dimension of experience, as well as to provide a solid theoretical ground for developing a method of emotion representation in artificial systems which will be proposed in chapter 4.

As argued earlier the mainstream definitions of knowledge that embrace only the objective dimension of conscious experience, i.e. facts about external reality encapsulated in intentional contents that have direction-of-fit, are not sufficient to embrace the complete experience. There is a need for systems that are able to represent the subjective dimension of experience that reach beyond objective knowledge which validity can be checked by mere matching the known facts with the state of world in the third personal terms.

In this section we will call evidence to the claim that subjective components of experience such as feelings and emotions complement human knowledge, and are

involved in cognitive and social processes, and are thus fundamental to purposeful behaviour, for which reason should not be taken for granted during conceptualisation or design of information systems that aim at representing human knowledge/experience, emulating human rationality or are constructed with the purpose to study or support human decision-making.

3.3.1 How emotions and feelings fit in conscious experience

Let us start by pointing out that starting in 1980 and early 1990s we have been witnessing a significant shift in perceiving the role of emotion in human rational behaviour and intelligence in science.

The significance of emotional dimension of human consciousness and action has not been the primary focus of mainstream scientific discourse in the past century however. Only in the recent three decades the emotional dimension of human existence has made its way to the central interest of behavioural, social and cognitive sciences. Still, the apparent supremacy of 'the rational' and 'behavioural' over 'the emotional' in the scientific and intellectual discourse over past decades is less evident once one takes an in-depth look at both ancient and modern classics in philosophical thought at which point it becomes clear the emotional dimension always played an important role in disentangling the human nature. Nevertheless these are only the recent findings brought by modern cognitive neuroscience, equipped with latest brain scanning technologies, that constitute solid evidence for placing parts of the brain responsible for emotional management in the centre of cognition, action and overall life management. Tribute is due to classic "armchair" theorists that in clearance of their minds managed to anticipate the results now visible on MRI scanners.

The evolution of emotion theories from modern-classical James's account of emotion as perception (James, 1884), Cannon-Bard theory (Cannon, 1927), Lazarus (1991) cognitive approach, to recent theory of Ledoux (2000), which says that emotions are result of both physiological reaction of brain and body, and/or mental interpretations related to a given situation claiming that there are different brain sub-systems for different types of emotions, clearly shows that emotions should be seen holistically as neurocognitive phenomena entirely depending on the processes in the brain. Furthermore it reminds of the necessity of revision of old theories of emotion and taxonomies to make them compatible with resent evidence from brain science, as many of these outdated theories and systematizations are unfortunately still being used and appreciated outside the domain of neurocognitive sciences, in particular by information scientists (please see Chapter 5 for evidence).

Contemporary psychology defines emotion in general as negative or positive reaction to a perceived or recalled object, event or circumstance accompanied by a subjective feeling (Damasio, 1999). Damasio based on the functional neuroanatomy of central nervous system (CNS) suggests to distinguish between emotions and feelings

noting that their essence is different. According to Damasio emotions are a sort of biological programmes of actions carried out by the body (internally and externally, like movement, facial expression, changes in viscera, body temperature, internal secretion of different substances, etc.) merely complemented by a cognitive program involving ideas and modes of cognition. The way emotions are triggered is beyond control of consciousness and precedes the conscious appreciation of stimuli and feeling of the very emotion activated. In other words an emotion is a set of programmed actions triggered to a large extent unconsciously, meanwhile feelings of emotion are the appreciations of the changes in the body caused by the emotion. So feelings are passive images of actions activated by the emotion and recorded in the form of brain maps, and thus memorized in some part at least (Damasio, 2010). The emotion-feeling feedback loop is well explained by below quote from Damasio:

> "Seen from a neural perspective, the emotion-feeling cycle begins in the brain, with the perception and appraisal of a stimulus potentially capable of causing an emotion and the subsequent triggering of an emotion. The process then spreads elsewhere in the brain and in the body proper, building up the emotional state. In closing, the process returns to the brain for the feeling part of the cycle, although the return involves brain regions different from those in which it all started." (Damasio, 2010)

Importantly in their core, i.e. in the basic structure of the action programme, emotions are unlearned, automated and genetically determined, and they only can be modulated with willpower to a certain degree. There can be some plasticity in these programmes as far as the triggering stimuli is concerned of course, as different people fear and get angry about different things, also some external expressions can be hold under control, still the lion part of the inner processes develop automatically beyond the control of volition. So although ultimately it is the feeling that is what brain maps and remembers as the end product of emotion, the way how emotions function determines the feeling, so understanding the emotions lets us understand what is the feeling to it.

Furthermore Damasio (Damasio, 2010), distinguishes the so called *primordial feelings* which reflect the current state of the body along varied dimensions, e.g. pain and pleasure, originating from the brain stem, therefore evolutionary older, part of the brain rather the the younger cerebral cortex. Interestingly according to Damasio but somewhat similarly also to other neuroscientists such as Jaak Panksepp (Panksepp, 2005) and Derek Denton (Denton, 2005) the primordial feelings constitute the basis for *a self*, referred to as the protoself or primordial self. This is a critical hypothesis, as it proposes that self is in fact build up with feelings which are the image of a leaving creature captured by its brain. Therefore the famous Cartesian maxim "Cogito, ergo sum" should be replaced with a more Spinozist "Sentio, ergo sum". The

following two quotes from Damasio explain the mechanism of brain-to-body mapping and its character in neurobiological terms.

> "The body-to-brain signalling (...) does not deal merely with the representation [in the brain] of quantities of certain molecules or degrees of smooth muscle contraction [bodily states]. (...) But there is, side by side, a qualitative aspect to the results of the transmission. The state of the body is felt to be in some variation of pleasure or pain, of relaxation or tension(...)." (Damasio, 2010, p. 97)

> "(...) the regions [of the brain] that receive body-to-brain signalling respond, in turn, by altering the ongoing state of the body. I envision these responses as initiating a tight two-way, resonant loop between body states and brain states. The brain mapping of the body state and the actual body state are never far apart. Their border is blurred. They become virtually fused." (Damasio, 2010, p. 100)

The key conclusion from the above quoted insights on emotion supported by neurological evidence is that whatever mental state is experienced by a conscious mind it has an inseparable affective part. If there appears an intentional content in the mind that we could qualify as knowledge (a proposition for instance) it is accompanied by a primordial feeling reflecting the affective state of the entire body. Importantly, as we know from earlier sections of this chapter, these two components cannot appear in consciousness as two independent states occurring in parallel but rather they are amalgamated in a unified field of consciousness.

It is the right moment to recall that in the section on qualia we have concluded that, in line with the representationalist account of qualia, the nature of the subjective component of experience is also intentionalistic in nature. The difference is that subjective qualitative intentionalistic contents that correspond to qualia are directed towards *within* the mindbody. Such an account has been inspired by and supported by the account of protoself and primordial feelings as formulated above by Damasio. Now, in accordance with both the theory of unified field of consciousness as well as the concept of "two way resonant loop between body states and brain states" proposed above by Damasio it must be that both types of intentionalistic contents of mind, those directed outward and inward, must be unified, are the two sides of the same coin.

This entails that each intentional content which can be qualified as knowledge, or image captured by senses and recorded in the brain, or each abstract thought that is a product of manipulation of images mapped in the past is affectively marked. And this marking plays and important role in cognition and behaviour. The role is triple at least: (i) the affective marking governs the selectivity of conscious attention; (ii)

impacts behaviour both reaction as well as deliberative voluntary action; (iii) alters memory processes.

The first role is underwritten by a somatic marker hypothesis formulated by Damasio, according to which images captured by the mind by applying its intentional capabilities are somatically marked (Damasio, Everitt, and Bishop, 1996). This hypothesis served him to explain how the selection and ordering of the abundant quantity of images captured by the mind is managed. The hypothesis proposes that the marker signalling influences the response of the organism to stimuli. In the process of self information the brain automatically manages external signals in a way that it tags these with somatic markers that represent the body state which prepares the organism for instant response in case this is necessary. The markers are termed somatic because they relate to body-state structure and regulation represented in the brain. This mechanism is subject to learning so that the tagging is not fixed but subject to adaptation along with experiences, some kind of learning algorithm seems to be applied as although the marking process may be open to consciousness it is clearly not under its control. The practical significance of this process for conscious and unconscious life management is critical.

The somatic marker mechanism according to Damasio is used not only to orchestrate reactions but also applies to attention selection. The somatic markers are emotional states, not necessarily a fully fledged emotions but some sort of feelings, that highlight important images captured by the brain and bring them thus to attention of consciousness. So the influence of these emotional markers is double, first they impact the way conscious attention rations the masses of sensory data or all sorts (visionary, auditory, tactual, etc.) bringing to the attention of consciousness only those that according to the marking mechanism are important, from evolutionary perspective critical for survival and life management, secondly they influence the response.

This capability is underpinned by non-conscious mechanism referred to as preattentive processing identified in neurocognitive studies involving analysis of reaction time to affectively heavy stimuli such as happiness/anger facial expressions, or images of spiders and snakes which has showed that processing of emotional stimuli in non-consciousness occurs before the operation of selective attention which results in enhanced detection of stimulus with affective value (Dolan, 2002). The significance of emotion and feeling to behaviour is well emphasised in Dolan:

> "Emotion provides the principal currency [ascribes value to events] in human relationships as well as the motivational force for what is best and worst in human behaviour. Emotion exerts a powerful influence on

> reason and, in ways neither understood nor systematically researched, contributes, to the fixation of belief."[7] (Dolan, 2002, p.1191)

In the view of what has been said so far, while considering the influence of feelings and emotions on behaviour it must be made clear that different types of affective states have different consequences for organisms behaviour. The number of these types is certainly not mapped completely which poses a significant challenge to any effort aimed at emotion/feeling-to-behaviour systematization. The basic distinction which must be made here is that between how emotions and feelings impact behaviour.

As mentioned earlier emotions are largely genetically determined programmes of responses to a stimulus. So emotional response, although modulated with conscious perception, develops largely autonomously, according to a relatively well studied patterns. For this reasons emotions in this sense fall beyond behaviours dictated by voluntary, deliberative decision making. On the other hand feelings, which affectively mark intentional conscious contents of mind must have far reaching consequences on deliberation processes. The impact of feelings, serving as the value meter of intentional states, on deliberation is manifold. Firstly, (i) the reasons for actions are affected, as feelings exercise motivational forces, dictating what is wanted and what is unwanted by the organism. Secondly, (ii) the feelings, in line with the above quoted account by Dolan, influence our believes, so they affect the content of the intentional states, which are processed in the deliberation process. The way how they precisely do it according to Dolan remains a research challenge. Some preliminary hypothesis however can be formulated and will be discussed in the next section.

The possible ways of how feelings influence reasons for actions is interesting in particular. It is postulated in the somatic marker hypothesis earlier discussed that the prefrontal cortex that is a region responsible for volitional behaviour provides, during deliberation, access to feeling states related to similar decision situations confronted in the past. Consequently the past feelings advocate for or against taking action toward or away from deliberated action options.

This also explains the neurological mechanism behind the decision heuristic studied by Klein (Klein, 1999) - Recognition-Primed Decision model - that emphasized the role of intuitions, previous experience and "gut feeling" in taking even very serious decisions. The heuristics described by Klein, based on the psychological studies of fireman behaviour in action, involved first picking up the top-of-mind solutions scenario dictated instinctively, then an imaginary projection of its implementation. Given a chosen solution candidate fails this mental simulation test another one is tested in the similar way until some solution candidate passes it. In this mental process of solution evaluation feelings must play important role, if the first scenario is picked up based on the "gut-feeling" likely the evaluation is done in the same way.

[7] It is evident from the context of this statement that Dolan while speaking about emotions as the principal currency must have meant the feelings given the emotion-feelings distinction proposed by Damasio."

Naturally this heuristic has nothing to do with blind guessing but is underwritten by a precise mechanism implemented in neuronal circuits that process affectively marked images recalled from memory, engaging prefrontal cortex and relying on brain capacity to mentally simulate outcomes of prospective actions. In this simulation feeling states corresponding to imagined outcomes of implementing a tested solution are evaluated. Consequently the solution that resulted in most positive affective value of imagined outcome is put in action. This mechanism most likely stands behind many heuristic that fall under the umbrella of what is vaguely referred to as intuitive behaviour.

Finally, (iii) emotion and feelings impact profoundly memory processes and learning. Affective marking of stimuli served from the evolutionary perspective pragmatic purposes: fast and accurate predictions regarding occurrences important from the viewpoint of survival, and adopting relevant response in case of re-occurrence in the future. Affective marking enhances memory accordingly, as to allow organism recall and pay attention to important events. The critical role in this process is played by amygdala as research evidence proves that patients with bilateral damage of amygdala nuclei do not have the advantage of enhanced memory for affectively marked images (Cahill et al., 1996). What is more, there exist neurochemical mechanisms which augment memory of affectively marked images (McGaugh, 2000).

Again it is important to distinguish the memory processes that function at the level of feelings from those affected by emotion programmes as well as on whether these operate on the conscious or unconscious level, as fMRI-aided research shows that consciously and unconsciously projected emotional stimuli are handled by different neural networks in the brain, different nuclei of amygdala in particular (Kandel, 2006, p. 219). More detailed discussion on the impact of emotion on memory and consequently forgetting is covered in section about temporal dimension of consciousness.

An important conclusion from this deliberations and somatic marker hypothesis is that intentional contents of mind are emotionally marked. A parallel could be made to semantic tagging of information content. Somatic markers are a sort of biological metadata for knowledge proper: the intentional contents represented in language. Once again we see that knowledge limited to intentional states expressed with language cannot embrace this meta component which apparently is vital for orchestrating the behaviour of any organism, certainly it is the case of a human being.

The next important question, once we have answered to what stands behind the qualitative subjective component of experience, is what qualities we are talking about in first place. In other words if affect colours experience what colours are involved. To answer this question we would have to refer to kinds and properties of emotions and feelings and eventually map these onto the corresponding behavioural consequences.

Although complete mapping of affective states to behavioural patterns would call for a time and effort intensive research agenda, the starting point for achieving this is the basic understanding of the nature of emotions and feelings. In chapter 4 we will provide a nutshell overview of types of affective states and their most important qualities, as well as we will single out some examples of how these states could influence voluntary action. Here let us conclude we an overall mapping between affective and experiential states at a more general, meta level.

Both emotional feeling states and experiences are subjective and are characterised by valence (positive-neutral-negative) and intensity (high-low). We propose that experience are higher level phenomena that embrace knowledge and affective states which provide experiential states with above mentioned properties (valence and intensity). So the qualitative value of experience is *inherited* so to speak from the feeling states originating at the primordial low=level brain processes.

Consequently experiences can be assessed in a qualitative way. Experience can be positive - result in feeling of pleasure, or negative - result in lack of pleasure or pain. Furthermore experiences could be ordered according individual preference: one can say that experience x is better, worse or equals to another experience y. As discussed earlier, qualitative value of experience (quale) is representational property of external objects or other stimulative events, of whichever nature, causing experience x and y. Evidently, as experience is subjective the same stimulative entity can result in experiences of different quality depending on the subject, or circumstances (varying in time and space), however should not vary between two identical stimulative events, although such identical events could be extremely rare of impossible in real life, being theoretically possible.

The above properties make experience assessment and comparisons extremely difficult and complex, though psychological literature provides some examples of both theoretical and empirical scientific endeavours challenging this problem. Maslow developed a notion of peak experience, which he defined as the moment of highest happiness and related it to self-actualisation (Maslow, 1971). Similarly, the optimal experience was described by a Hungarian psychologist Csikszentmihalyi, who introduced a concept of flow (Csikszentmihalyi, 1975), the moment of top experience when one is confronted with a demanding challenge, still attainable with one's own capacities, and deeply enjoys the moment of stretching intellectual capabilities, and thus learning and increasing self-esteem. We will investigate the qualitative nature of experience inherited from affective states further in chapter 4.

3.3.2 Limitations of language in expressing affective states

The problem with talking about emotions and subjective feelings is that we lack language for their precise description. Traditionally entities which have been described as "the rational" linguistically pertained to the realm of science and philosophy

meanwhile "the emotional" pertained to the realm of arts. This is exactly because these two realms are governed by different narratives, the former by the inter-subjective and objective reality and the latter by primarily subjective, intra-subjective reality. Speech acts are forms of institutional intentional states therefore linguistic components have corresponding parts of objective reality that they represent. Everyone knows what a "wheel" is because one can point at an object that she and others can see and say pointing a finger "This is a wheel". This cannot be done with emotions because the emotions are strictly subjective. Put in philosophy of mind terms this represents the impossibility of determining the *direction-of-fit*. Of course one may verify the statement "A is angry" by matching particular bodily expressions to the emotion of anger, still how doest the anger feel to A, what are its variations, intensity, etc. is hard.

It is important to notice that when it comes to subjective mental phenomena the language that is the carrier of subjectively felt mental states, and the observed behaviour, remain the only proven, yet imperfect, tools for interpersonal comparison of these states. This is why the lack of proper terminology, or/and terminological disarray, as well as language imperfections are an important obstacle for understanding of internal subjective mental phenomena that give flavour and meaning to human experience. Likewise, methodological disputes on whether observed behaviours can be taken as solid basis for inter-subjective comparison of internal mental states that apparently caused that behaviour, as in the above used example of the angry A build up additional hurdles. Therefore it is important to reflect on the limitations of language we face in the study of human emotion and experience at large and consider possible solutions or ways to go about it.

First let us notice that language is a third-personal, inter-subjective construct which serves people to communicate, i.e. share contents of their intentional minds and act in groups. Searle while writing on intentionality (Searle, 1983) highlights two important properties of language: he points out that speech acts are a type of human action, and thus become the part of social reality, speaking is behaving and language is both behavioural and social phenomena, and forms of intentionality underlying language are social forms. Needless to say language also serves other important purposes, particularly it aids mental representation and abstract manipulation with intentional contents, as well as building a "narrative" part of self, though it seems not its primordial role. So the first limitation of any language in expressing affective content stems from sole fact that language is intrinsically inter-subjective, social phenomena meanwhile feelings are privately, subjectively appreciated.

Jaak Panksepp points out that it is necessary to tell apart the propositional speech from the affective contents it carries in form of intonation variations. Speaking form the evolutionary perspective, he observes, that early vocal signs which preceded fully developed propositional speech apparently served the communication of emotional expressions, these however have only weak links with propositional speech which

has evolved much later and most likely have been developed as an effective way of encoding the relationships among external events. Even in today's human beings the voice intonation and social colouring of utterances is handled by a different hemisphere of the brain (right) than it is the case for propositional speech emerging from the left hemisphere (Panksepp, 2005). This also explains why music with its richness of intonations and timbre is such an effective medium for affective communication, albeit lacking words.

Furthermore, it is increasingly believed that language is not the intrinsic form of mind intentionality and other non-linguistic forms of intentionality should be studied after the dominance of philosophy of language (Searle, 2008). What is more, contrary to the Chomsky's universal grammar theory and other linguistic theories proposing that some of the properties of language are hard-wired into the brain, and encoded genetically (Chomsky, 1965) contemporary brain research provides sound arguments to the claim that acquisition of language is dependant on emotions, suggesting that affect lies at the basis of mind's intentional capacity. Greenspan and Shanker have formulated the following hypothesis, which resulted from their extensive therapeutic work with autistic children:

> "Our ability to form symbols, which enables us to represent our world and reason about i – and all the great intellectual accomplishments that build upon this – has an unexpected origin. In order to develop symbols, we must transform our basic emotions into a series of succeedingly more complex emotional signals. This human capacity to exchange emotional signals with each other begins in early life during an unusually long practice period and leads to symbols, language, abstract thinking, and a variety of complex emotional and social skills that enable social groups to function. the exchange of emotional signals may also play a critical role in the development of the brain, especially that of the higher cortical centres dealing with language and thinking, the prefrontal cortex dealing with planning and problem solving." (Greenspan and Shanker, 2004, p. 17)

The above hypothesis is indeed ground breaking. It supports our claim that affect stands behind meaning. Single words and utterances carrying intentional contents have meaning in semantic terms, which corresponds to the contents of the intentional mind state as well as a meaning in a more subjective, private dimension, which is provided by remembered, accumulative affective value corresponding to these contents.

Panksepp makes it explicit that substantive understanding of emotions cannot be generated by sole means of using the language. To talk about emotions and feelings in a precise, scientific terms, he proposes, neural criteria must be taken into account. Astonishingly, at the same time topically for scientific disputes, despite common agreements on basic affective processes engineered in the brain by evolution,

terminological disputes and disarray continues and there is little agreement on how affective functions should be labelled, described and discussed in the scientific discourse.

The ways out of this "linguistic" deadlock include: (i) rationing the debate to observable behaviour as behaviourists postulated, (ii) pushing the language to its limits, using to the largest extent possible analogies, metaphors and both scientific jargon and vernacular terms to build viable and understandable concepts, in a narrative, descriptive manner, (iii) to define a formalism for standardized description of feelings including feelings of emotions, believing that these are interpersonally comparable.

Behaviourism, especially in its extremes, has been widely rejected as unsatisfactory narrow. We favour a mixture of (ii) and (iii) or better said (iii) elaborated via (ii). We will propose in chapter 4 that feelings, including feelings of emotion, in the same fashion as other complex yet understandable and reducible phenomena that involve subjectivity could be represented with a crafted symbolic system that builds on the inter-subjectively comparable indicators, such as activation or inhibition of relevant brain circuits. This will build on the analogy with systematic, standardized representations of reach colour patterns such as RGB or CMYK although the analogy is not fully entitled as colour schemes describe colour objective properties not colour qualia. Let the below quote from Panksepp serve both as the explanation of this subtle difference as well as the conclusion of this subsection.

> "The basic emotional affects are primary brain/mind processes, similar to seeing a color. One can use a word, like "red", as a label for a color, but this word does not explain the experience of seeing red. If someone is blind, the word "red" is meaningless. In order to explain seeing red, one must discover the neurophysiological and neurochemical causes of visual experience. Similarly, one cannot use words to explain primary-process raw emotions. Words can only be used as second-order symbols to discuss affective experiences, but they do not adequately capture the fundamental causes of feelings." (Panksepp, 2008, p. 49)

3.4 Experience as self-information and its role in purposeful behaviour

This section shall be dedicated to the subject of knowledge and experience significance to purposeful behaviour. Zeleny took and interesting perspective on knowledge saying that:

> "Knowledge is purposeful coordination of action. Achieving intended purpose is the sole proof or demonstration of knowledge."(Zeleny, 2002)

Although this definition seem to manifest philosophical behaviourism which we reject it well illustrates the very purpose of knowledge, its pivotal role in purposeful behaviour. From the pragmatic standpoint that we favour only this aspects of knowledge that are relevant to behaviour shall attract our attention, which is why we decided to include this section in the final parts of this chapter. This section will also provide a prologue to chapter 5 in which we will discuss possible applications of the conceptual frameworks for experience representation, which will be elaborated under in the following chapter 4 to modelling rational behaviour.

The theory of unified field of consciousness suggests that separating emotions from reason does not make much sense. Rationality, or rational behaviour is a function of both conscious and unconscious processes where emotional reactions and actions are on an equal footing with cognitive and deliberative behavioural responses composing together a complex stimuli-action management system. In the same way as dualism proposed by Descartes is being abandoned for creating false mind-body problem the dualistic approach to human rationality that opposes rational judgement to emotions must be left behind. Evidently there is a need for a new coherent theory of human action that unifies cognitive, emotional and deliberative dimensions, which is in line with both intuitive, introspective theoretic philosophical insights and empirical evidence coming from brain science. This must be possible as indeed the source of it all is one. Such a theory would without a doubt provide a significant push forward for the AI methods and techniques, in particular would bring closer the vision of building up accurate systems that emulate human behaviour.

3.4.1 Practical reason and affect

It is almost trivial to state that brainmind is an organ responsible for life management, i.e. managing organism's behaviour. Behaviour is understood widely here, spanning from physical bodily movements to acts of thought or speech acts. Interestingly a vast majority of the processes in the brain lead to some kind of movement, limbs movements, speech apparatus movements, heart beats, handwriting, facial expressions, gestures, etc. Daniel Wolpert suggests that complex central nervous systems including elaborate brains has evolved due to the requirement of adaptable complex movement. These assumptions are based, among others, on studies of a simple animal that after a period in its life where it moves eventually it settles and first thing it does after settling is to digest its CNS for food, apparently not needing it any more. There are more other indications for such an account, including trees that are complex biological beings, yet do not have the capacity to move and lack CNS. There is no point in thinking process if there would be no future/potential expression in action

Zeleny binds knowledge with action, inextricably linked with movement, in a reflection provoking way:

> "Knowledge is purposeful coordination of action. Achieving intended purpose is the sole proof or demonstration of knowledge."(Zeleny, 2002)

Although this statement may appear too far-fetched, it could be rephrased to be more definite: knowledge manifests itself in purposeful coordination of action.

Damasio depicts brain as a great cartographer (Damasio, 2010) with this metaphor he aptly wraps up the very purpose and at the same time the nature of the mind. Brains build up maps that are needed for life management, which is the essence and the eternal goal for each organism including human being. Therefore the reason for having brain, thoughts, mind and at the same time their main occupation is self-informing, which is the necessary condition for conducting action, behaving, in particular decision taking within purposeful behaviour concentrated on life management. Such behaviour is immanent for every life forms starting from single cells ending with highly complex minded and conscious beings like humans.

It is obvious that knowledge is the basis for human behaviour. Any theory of human decision making would include an element of deliberation, in which an agent considers means and ends. Even strict behaviourists accept that there must be some thought process going on in the black box, though they find it irrelevant, unlike consequences of the behaviour. The popular choice theory proposed by Simon and Newell in which decision making is framed as information processing is a distinct example (Newell and Simon, 1961; Simon, 1978). Still an important question remains to be addressed, whether the qualitative dimension of knowledge, the missing component of conscious experience makes any difference to the decision making process. So we ask here if the fact that conscious experience has a subjective feeling matters for practical reason, and if so, how exactly? To address this question we first have to consider whether and how, mental states that are not physical phenomena can have any causal effects on physical reality, which constitutes the so called *mental causation* problem encapsulated in *the mind-body* problem.

Searle provides an answer to it, which stems directly from his account of the mind-and-body problem referred to as the *biological naturalism*, discussed earlier already. According to Searle mental causation is possible because conscious states are real features of the real world despite being "mental". Importantly biological naturalism rejects dualism, i.e. the separation of mental and physical phenomena treating them solely as two different narratives or two descriptions at different levels of the very same thing, "one complete system". Still the distinguishing element of Searle's view is that he also rejects reductionism of mental states to physical brain states (Searle, 2004, p. 78) keeping the belief that qualitative feeling of mental states is not illusionary but real. According to Searle the mental causation is possible just because the basic assumptions of dualism that led to the problem itself are false and should be abandoned. When there is no mind-body dualism there is no problem with mental causation. So it seems that rejecting the dualism is the only way how mental

causation, causation at large and specifically the causality in practical reason can be accepted.

Now, the critical point in this line of thought is to realise that mental causation which we encounter in purposeful behaviour, voluntary action in other words, is significantly different from any other causation we experience in the physical world. This point is made very clearly by Searle and explained in detail in his book "Rationality in action" (Searle, 2001). The difference in question is that unlike in the case of causality we experience in the external world which must be stemming, in Humean terms, from the presupposed uniformity of nature guaranteed by causality and causation that is given and nothing can be done to eliminate or question it, in case of the intentional action of a minded creature there is a causational gap. This gap is the gap of free will. We are discussing the problem of fee will in the next subsection in more detail as it is central to the problem of practical reason.

Searle answering the scepticism of Humean problem of induction notices that the existence of the causality in the external world can be extended from the causality that we experience in our waking consciousness (Searle, 2004, p. 144). As we experience in our mental states the cause and effect relation we can assume that the same causality feature functions in the external world. Although this stance could be questioned on the basis of the possible erroneous or illusionary character of our perceptual experience at large, the perception of causality should not be less unlikely than any other perceptual experience.

If indeed there is the gap of free will there is no other way than accepting the following scenarios: (i) the gap of free will is real, completely indeterministic and unexplainable in causal terms, (ii) the gap of the free will exists, is real, but we still cannot explain how it works, it may be deterministic may be not or may be of some other yet not uncovered nature, we may learn one day, (iii) there is no gap of free will, we will soon learn it from the brain studies which will show that free will is an illusion and so called free action is fully determined by the unconscious processes in the brain.

Searle is a strong supporter of the first claim. Here let us propose that the second alternative seems the most likely scenario but importantly let us observe that although indeed under the current state of play the gap of free will seems to be real, it can be, at least partly explained by looking into the realm of the subjective experience. If we are ever going to explain the voluntary action of the mental in causal terms we cannot ignore the subjective feeling of conscious experience. Searle points at the minded self as the cause of the gap, meanwhile it is evident that emotions and subjective dimension of experience is the place to look for the self.

Consequently if it is ever to be possible to emulate purposeful behaviour of minded creatures in artificial systems it can only be done by proper representation of the subjective, affective dimension of the experience and its causal mapping on purposeful behavioural patterns of these minded creatures, humans in particular, or

at least such a representation shall be seen as the necessary milestone along the way. However if Searle is right, that the first assumption is correct than most likely we will never be able to represent human purposeful behaviour in artificial systems.

So the first important moment when affective, qualitative dimension of experience plays in the decision making process is in who it fills in the gap of free will, given Searle is right that the minded self is responsible for the gap and Damasio is right that self is most likely build up from primordial feelings.

Damasio notices that one of the main capacities of highly developed brains, such as human and primates have, a capacity which makes complex life management of an organism possible is creation of maps. These are mental maps, representations of the body and external world in the brain, which are the way the brain that is responsible for life management "informs itself". This map making is done while the body of and organism changes and as it interacts with the environment. One type of this maps are sensory images that can be manipulated in the consciousness and to which reasoning can be applied. (Damasio, 2010, p. 63). This, naturally, has a profound impact on how the behaviour of an organism, in particular the voluntary behaviour, i.e. decision making, is performed. The maps created by the brain are the very basis for decision making and the realisation that these maps do not merely concern the external environment sensed by the traditional five senses: hearing, sight, touch, smell, and taste but also maps of internally subjectively and qualitatively experienced internal states sheds new light on how rational behaviour is perceived. Especially because the boundary between these maps is difficult to spot and very likely the working human brain spontaneously (or by a fixed patterns that we still do not know to date) associates elements of one map with the other, which very much likely is the source of human and some animal creativity. The mechanisms that are responsible for this process will stand for what Hofstadter called the creative analogies, the very essence of biological intelligence.

The interdependence between the realm of objective thought and subjective bodily feelings is figuratively yet aptly depicted in the below quote from Damasio:

> "Body and the brain are engaged in a continuous interactive dance. Thoughts implemented in the brain can induce emotional states that are implemented in the body, while the body can change the brain's landscape and thus the substrate for thoughts. The brain states, which correspond to certain mental states, cause particular body stated to occur; body states are then mapped in the brain and incorporated into the ongoing mental states."(Damasio, 2010, p. 96)

There is enough evidence from neuroscience to claim that emotions play a central role in human and animal decision making (Damasio, 1994; LeDoux, 2000). Based on this evidence let us postulate that any intentional state, which is as explained earlier reflections of external world in human mind, has certain emotional value, which is

characterised by valence (positive or negative) and intensity (arousal level). The emotional value of intentional states that build up the total reason for action does influence the choice made by the agent.

Emotion feelings are central to trigger decision and actions (Damasio, 1994) as well as in deciding which option to choose. Saltzman and Newsome (Salzman and Newsome, 1994) research on neural mechanisms for forming perceptual decision shows that a deciding brain presents increased neuronal activity in certain of its parts as if it 'accumulated' simultaneously arguments for different available options, the choice is made the moment one options 'prevails' which is manifested by visibly strongest neuronal activity in one of the parts taking part in this 'neuronal dispute'. This suggest that decisions are made based on conscious evaluation of available options by internal collection of arguments for and against available alternative, which is biased by emotions.

So apart from the role in filling up the gap of the free will subjective feelings provide also currency for mental deliberation together and inextricably with outward intentional contents of mind, i.e. knowledge.

3.4.2 The central problem of the freedom of will

Searle (Searle, 2001) rightly points out that in order for an action to be a genuinely free none of the reasons for action can be causally sufficient. Truly voluntary actions cannot have psychological casual conditions which determine them. Thereby Searle talks about a gap that results in what we know as free will. In order to perform a voluntary action an agent must exercise its free will while making a decision based on the given reasons for actions. Such an account of fundamental principals of natural agent behaviour poses a real challenge for AI. There are basically two ways to go around this problem. Firstly, it could be assumed that free will is an illusion as, despite we do not feel it this way, actions are fully determined by the neurons in the brain or even deeper by the quantum states of particles building up our body. Secondly, we could admit the gap and implement some kind of free will prosthesis based on some kind of random process. Neither approach seems fully convincing.

Let us accept that human agents do have a genuine choice, importantly however their choice is *bounded* by the fact that they are intrinsically compelled to make decisions in first place. Long-term and persistent decision avoidance qualifies as mental disorder and must doubtlessly lead an individual to deterioration of life quality, general dismay, and likely death. Similar pragmatic argument is raised by Bayesian rationalists in favour of classical rationality model which presupposes that a rational agent is guided by maximising its subjective expected utility (Edwards et al, 2007). But the difference is important as we do not use this argument in favour of any particular decision model but to support the claim that a rational agent cannot avoid decision-making. In other words a human agent whenever confronted with

a decision will have to respond with some action, either would go for one of the available alternatives or abstain from acting, which is always one of the available options. As time cannot be stopped refraining from action is a sort of action.

An interesting question that scientists struggle to answer is what is the source of free will, how it is implemented in the brain and how could it be explained theoretically or empirically. One hypothesis suggests that the source of free will should be sought in some deep, low-level (subatomic) indeterministic (quantum) processes, which are the only indeterministic processes we know in physical world, which has been proposed by some philosophers including Searle (Searle, 2008). Another one points at system-level processes that can give birth to emergent, epiphenomenal properties of a complex systems which a biological brain doubtlessly is. Yet another one suggests that the source of free will could be found in a more generic capability of the brain which in literature is referred to as *neuroplasticity*.

Brain is an incredibly flexible organ, clinical cases of patents with brain lesions caused by strokes or injuries with similar effects show that brain can relocate its functional specialisation areas to compensate for the cognitive losses caused by the damage. This capability particularly manifests during a relatively short period of time, 2-3 weeks, immediately after the damage, as rehabilitation proves much more effective during this early days after an incident. The sole fact that behavioural therapy and rehabilitation prove to impact the recovery process is a blunt evidence that brain has a self regulatory mechanisms within which we as conscious beings can identify by introspective experience the free will as a part of this self-regulatory system. This self-regulatory system is likely therefore to be responsible for such crucial brain capabilities as: learning, motivation and free action, and also indirectly bio-cultural co-evolution.

Likewise, the brain should be capable of significant modifications as a result of internally directed focus of attention: self-reflection, mindful meditation. Based on the studies on mindful meditation recalled in earlier sections it can be assumed that human brain is flexible enough to change the way it functions in course of conscious mental process which goes beyond typical capabilities of a normal fully functional healthy brain. It is still uncertain, but very likely, that many of the brain processes that have been believed to be hard-wired, such as fear-governing brain circuits could be altered with mindful meditation and mental practice. This could allow for a volitional, mental control of fear and other emotional reactions that are normally, physiologically beyond the control of consciousness.

This brings us however to a visible contradiction between two opposing accounts on the free will problem. One account proposes that the free will is real and suggest that it could embrace more and more mental processes via inward directed volitional control or mindful meditation. Another favours reductionalist stance that free will is an illusion, can be fully explained by lower level processes in the brain and even more these processes are governed in the unconscious. The supporters of the latter

view often use evidence provided by recent discoveries of neuroscience as important arguments in the discussion. Still, the evidence from brain research has so far provided inconclusive arguments to the philosophical discussion on the freedom of will.

We know without doubt that to some extent free choice is determined by unconscious processes, or at least the brain unconsciously gets prepared so to say for the execution of of the voluntary action before the decision about the action reaches consciousness. This has been shown in the studies carried out by Libet (Libet, 1999), largely inspired by previous work in the area including Deecke (Deecke, Scheid, and Kornhuber, 1969), reporting that

> "Freely voluntary acts are preceded by a specific electrical change in the brain (the *readiness potential* - RP) that begins 550 ms before the act. Human subjects became aware of intention to act 350-400 ms *after* RP starts, but 200 ms. before the motor act. The volitional process is therefore initiated unconsciously. But the conscious function could still control the outcome; it can veto the act. Free will is therefore not excluded. These findings put constraints on views of how free will may operate; it would not initiate a voluntary act but it could control performance of the act."

Libet's account is thus as follows. Doubtlessly a voluntary conscious act is initiated unconsciously however there is a place for a free will in a form of rejecting/accepting mechanism for the action programme suggested by the unconscious processes. The interesting point on which Libet reflexes is whether the voluntary veto act is also unconsciously pre-determined. The latter hypothesis is reject which is explained as follows:

> "(...) one could consciously accept or reject the programme offered up by the whole array of preceding brain processes. The awareness of the decision to veto could be thought to require preceding unconscious processes, but the content of that awareness (the actual decision to veto) is a separate feature that need not have the same requirement."

Such an approach to free will as a "veto mechanism" has been labelled *free won't* by Hofstadter (Hofstadter, 1979).

The RP hypothesis suggests therefore that initiation of the freely voluntary act begins in the unconsciousness which already introduces a completely new perspective on the freedom of will. Notably, there is a qualitative difference between the freedom to veto and the freedom to suggest actions, an analogy to different models introducing separation of powers studied by political scientist inspires the imagination. The creativity process for instance under this account would fully rely on the unconscious processes, furthermore ingenuity of a free character of a voluntary act can be easily undermined. Importantly it has provided support argument to physical determinants

who somewhat prematurely judged that brain science proves that free will is an illusion. However the RP hypothesis has also been criticized. Searle for instance rejects Libet's account providing the following justification:

> "Libet's description lends itself to interpretation that the readiness potential marks the onset of the action. But that is not true. There are typically about 350 milliseconds between the readiness potential and the onset of the intention-in-action and another 200 milliseconds to the onset of the bodily movement. In any case, as far as we know from the available data, the occurrence of the readiness potential is not causally sufficient for the performance of the action. (...) It seems clearly premature to assume that the existence of the readiness potential shows in any sense that we do not have free will." (Searle, 2001, p. 291)

Although as the juxtaposition of the above quotes from Liebet and Searle shows, there is a bit of misunderstanding on the extend to which Liebet himself rejects the existence of free will, the reply by Searle illustrates that the sole proposal that because of RP the action is *initiated* unconsciously can be questioned.

More recent studies using fMRI technology by Soon et al. (Soon et al., 2008) suggest that predictions about the outcome of supposedly free choice can be made as early as up to 10 seconds before the decision reaches awareness. Unconscious determinants of free decision in the human brain discovered by Soon are located in the frontopolar and parietal cortices (higher-level areas of the brain), unlike it was in the case of Libet who investigated the SMA motor-related brain regions. The evidence provided by this studies suggest, in contrast with the previous studies, that unconscious brain processes not only determine an unspecific preparation of a response but also encode specifically how a subject is going to behave. Soon et al. reporting these results conclude:

> "A network of high-level control areas [frontopolar and paretal cortical areas] *can* begin to shape an upcoming decision long before it enters awareness."

We have emphasised the word *can* in the above quote to highlight the fact that even armed with apparently strong evidence showing that unconscious processes precede the voluntary decision making the scientists are extremely conscious about decreeing the non-existence of genuine freedom of will. The possible reasons for such a situation will be reflected in paragraphs to follow.

Most of the prominent brain scientists, like Damasio, Ledoux, Panksepp and philosophers like Metzinger, Searle believe that consciousness and free will, or free "won't" as Hofstadter aptly frames it, will be explained one day by brain science. As Searle (Searle, 2008) concludes there are potentially 2 possible directions of research adopting either a system approach to the problem or single brain area approach.

In the former case the studies would focus on examining different altered states of consciousness (sleep, coma, trance, consciousness disorders etc.) in the latter case the research would focus on close examination of phenomena such as Gestalt unity of consciousness. We propose yet another study that may provide some insight into the phenomena, namely what happens in the brain when we voluntarily let our brain accept an illusion (for instance a crafted visionary illusion) as true and after that "get back" to the acceptance that this is only an illusion. In other words which is the locus of control for putting ourselves in and out of an illusion.

It may be wrongly suggested that if some of the processes are unconscious these automatically means that these are not free. To us it can equally hold true that genuine freedom of will operates in the unconscious, and that the genuine character of the freedom of will is not conscious but unconscious, in which case we will never be able to verify it, as we will never have access to such knowledge consciously and inter-subjectively. Supporters of Wittgenstein's epistemological account and modern computer functionalists such as Dennett will say here that if we cannot say anything certain about that, this means that we are wasting time trying to make the point in first place, as that about which nothing can be said does not exists. Again we arrive here to a similar deadlock as in the debate on qualia and subjectivity of conscious experience.

Furthermore, the *readiness potential* hypothesis may be incomplete, it may be that it overlooks some important, still undiscovered, features of the free will process that all in one prompts misleading conclusions. Last but not least, regarding the dilemma from a metaphysical perspective, we will never have chance to learn if the knowledge on the basic-level brain processes on which the *readiness potential* hypothesis is proposed is valid, and if it is to be questioned by consequent research sometime in the future.

From the above standpoint it is very likely that the phenomenon of free will and its basic properties, whether genuine or illusionary, will never be fully explained. Emulation of free will in an artificial system could be achieved provisionally by applying probability calculus or if we come up with a better formal theory which would better simulate uncertainty as freedom rather than randomness.

Noteworthy, supposing that the gap of free will or free won't is real, it is operating, as mentioned earlier, within tight boundaries of largely natured motivational constraints and mortality. How the behaviour can be considered free at all if it is motivated by a purpose that is not set by the subject. Our drivers do not depend solely on us, the very fundamental drives such as desire to survive, desire of avoiding pain, sex drive, are beyond our control. Of course one may provide a handful of counterexamples like: people who commit suicide, monks that give up having offspring, sadomasochists, but for some reason these are behaviours that are extreme and rare, for some reason it is very unlikely that entire mankind commit suicide or decides not to have children or we simply cannot stand still and do nothing. Evidently

we have a deeply biologically rooted mechanisms that beyond the control of our conscious self dictate the rhythm of our lives. Two which extent these drivers are accessible and controllable by conscious self is still a largely unknown territory. Until this is clarified we must accept a pretty limited notion of freedom in our action, or perhaps full biological determinism. On the other hand until we prove it is possible to programme a rational artificial agent we will not be sure that freedom of will is indeed epiphenomenal and irrelevant to intelligence.

Although the problem of free will is central to understanding conscious experience, and consequently forms the foundation for providing a sound experience representation framework, it unfortunately appears to us as unsolvable dilemma at present. Noteworthy this assertion is not new whatsoever as thinkers and scientists have been confronting it constantly for as long as history of ideas reaches back, and despite the evident lack of clear cut answers the progress in science, including computer and information systems was possible and proved useful across many application areas, decision support and expert systems in particular. We therefore insist that despite the problem of free will can not be now solved efforts aimed at constructing *satisfactory* experience representation frameworks are justified, needed and can result in useful approximations applicable to solving real-life problems.

3.4.3 Emotions and feeling in rational behaviour

One in all, what we postulate is that we currently lack a complete theory of affective and rational action. This is difficult as mapping of all the affective and unconscious phenomena onto purposeful behaviour requires confronting complexities that perhaps go beyond capacities of our minds. Still, the building blocks for initiating construction of such a theory are there already and should be tried. We believe it should be started with identifying types of subjective feelings and mapping those onto intentional contents that should redefine knowledge and could pave the way for improved deliberation approximations.

It is important to explain here why our account is different from the early utilitarian theories, as a critical argument could be forwarded that what we propose is nothing but the revitalisation of the old utilitarian concepts, i.e. that people are driven by innate urge to avoid pain and seek for pleasure.

At a first glance this argument appears plausible however there is an important difference in the approach we adopt. The major difference lies in the complexity of emotional contents processing. First of all utilitarians focused solely on the pain and pleasure as the ultimate opposite subjective experiences. We suggest to perceive subjective experience as significantly richer palette of subjective feeling states which are either wanted or unwanted by an agent. To give a somewhat extreme example an agent may like to feel pain, or may dislike a feeling that is commonly liked by majority of other agents. Importantly there can be many types of feelings, at least

as many as there are fundamental emotional states, fear, disgust, shame, joy, etc. but potentially there can be a lot more of different shades and intensities of these emotional feeling states varying from a situation to another and having different sort of impact on agent's behaviour. Secondly, the outcome of the assessment process of the valence of the feeling state is much more complex, and we propose this complexity could be mapped by figuring out affective value of each piece of intentional contents and include them in the reasoning about the current, "real-time" affective state of the agent processing these contents while facing a particular decision situation.

Another important difference is the focus of utilitarians on the long-term perspective, a global optimum which is directing an individual to happiness and avoiding pain, ignoring the local optima, single decision in which not always an alternative which has the highest long-term pay-off is chosen, instead we point to the fact that the behaviour, and thus decision process by an agent is a "here and now" exercise, as no future nor past contents that do not appear in the present of the mind matter, so the currently simulated future pay-offs (importantly, simulation is made based on the remembered past experiences) only matters and only those that occur to, are recalled by an agent from his memory have any significance, not all those that are probable. More about different decision effects and their implication on the legacy behavioural theories will be presented and discussed in chapter 5.

Our account is also different from the standard approach to how emotionality is dealt with in so the called deliberative agents (please see chapter 5). Deliberative BDI agents are enhanced with emotions by enlarging their knowledge, beliefs to be more specific, set with propositions about their emotional states. However we do not accept such a solution as a satisfactory one because it confuses the state of being aware of an emotion with the affective state associated with any intentional state. This is supported and inspired by philosophical account by Donald Davidson:

> "According to Hume, 'reason is, and ought only to be the slave of the passions'. By this he seems to have meant that the passions (desires) supply the force that moves us to act, while reason (belief) merely directs this force. I doubt that desire can be distinguished from belief in this way; belief and desire seem equally to be causal conditions of action. But there is a sense in which desire can be said to be more basic conceptually. Desire is more basic in that if we know enough about a person's desires, we can work out what he believes, while the reverse does not hold." (Davidson, 2004, p. 26)

Desires are equal to beliefs as indeed a desire is in the above sense a "verbalised desire", i.e. in order for a proposition to be included in the deliberation an agent must have internally verbalize it and accept it by which he converts it into a belief. As a result an agent acquires a belief about his desire or emotional state. Moreover, as Davidson aptly puts it, a desire can be quantified according to its strength, which

under Bayesian choice theory equals to assigning subjective probability to it. Apart from desires, converted as just explained into explicit beliefs, there exist implicit affective states that directly influence behaviour, and these are not captured by the desires set in the BDI framework, neither by other instrumentalist rationality models because these currently are missing adequate forms of representation.

3.5 Affective bias in rational judgements – the empirical study

This section reports the results of an empirical study we have carried out in order to better understand the affective bias on rational judgements. The purpose of the study was twofold. Firstly, we wanted to empirically test that a strong affective stimulus would indeed bias a rational decision maker's judgments and secondly, to study the nature and strength of this effect in more detail, which would allow us to get a better grasp on the effect, much better than it would have been if we relied solely on the empirical data reported by other researchers in the subject literature.

In order to get a better understanding of the affective bias we have designed an experiment the structure and results of which are presented hereof. The structure of this report is aligned with the standards recommended by the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (Publications and Journal Article Reporting Standards, 2008).

3.5.1 Problem under investigation

The question we were trying to answer under the reported study was whether occurrence of stimuli that invokes strong affective response would influence rational judgement. In other words whether a properly catered affective stimulus can make individuals deviate from what is otherwise seen as a fair and balanced judgement of a selected phenomenon.

The relevance of emotional or affective bias in economic behaviour and rational decision making has been studied intensively by behavioural economists, and more recently by neuroeconomists. The overview of these studies has already been provided in the previous section. For further review of the topic please refer for instance to (Shafir and LeBoeuf, 2002; Loewenstein and Lerner, 2003; Rick and Loewenstein, 2008).

The influence of affect on rational decision making has been studied from different angles. One of the most commonly investigated problems is the influence of positive and negative affective states on risk assessment e.g. (Johnson and Tversky, 1983; Loewenstein et al., 2001). Knutson (Knutson et al., 2007) investigated the impact of emotion on consumer choice. We have decided to look into the influence of affect on rational judgement. Here by rational judgement we mean an assessment of quality

that is guided by balanced rational deliberation involving facts and cues that can be interpersonally appreciated. This does not presuppose that such judgement cannot be subjective. We have argued earlier that rational choice in most real life situations is subjective. For the purpose of this study it is sufficient to make use of the intuitive definition of rational judgment as fair assessment, that is one that follows the tenets of rationality.

More specifically, in our study we have focused on the negative affective state: *disgust* and its influence on the subjects's judgements about the quality of a selected wiki-type website, namely an open public Polish recipe service `www.wikikuchnia.org`.

Our primary hypothesis was that the occurrence of disgust deviates the assessment of the quality of wikikuchnia service and the user satisfaction from the service. So we assumed that the affect of disgust would lower the assessment scores for the service. Moreover we anticipated that the perceived tastiness of the dish being described on the chosen wikikuchnia entry would be influenced as well, and equally the disgusting stimuli would discourage people from recommending either the dish or the service to their friends.

In order to avoid response bias the respondents were misinformed about the actual purpose of the study. Furthermore, apart from the control group the study covered two study groups, one of which was exposed to an affective stimuli presented above the threshold for conscious perception (supraliminal) and the other below the threshold (subliminal).

3.5.2 Method

The study was carried out following the experimental testing regime with an experimental manipulation. A between-group design was applied with three groups involved: (i) the control group (Group C), (ii) the study 'supraliminal' group (Group A), and (iii) the study 'subliminal' group (Group B). The target participants where heavy Internet users, familiar with wiki-type services (e.g. Wikipedia) with basic computer and Web literacy. With this target set we have decided to target students or young people between 18-34 years old with invitations to participate in the study.[8] The samples were build and group selection made randomly. Invitations where sent out to 3 independent groups of students in Warsaw, Poznan and Lodz via e-mail and Facebook invitations. The targets were reached via tutors and students within the identified student communities. Personal and professional background of participants was assumed irrelevant. Participation in the study was anonymous, voluntary and unremunerated.

[8]Due to the selected questionnaire distribution method which relied on social media a few subjects from outside this target group participated in the study, however as the age was not a critical factor in the experiment these responses have not been removed from the sample

As for the research tool used in the study we have designed a questionnaire according to our own operationalization. The questionnaire was prepared with Google Forms web tool. There where free forms prepared, one per each group. Each participant received a link to the questionnaire corresponding to the relevant group. The questionnaire consisted of an introduction and the main part. The introduction covered the instructions for the respondents and the basic information about the study including the misinformation and manipulation, which will be explained in detail in the following subsection.

For the full questionnaire please refer to Appendix 1.

3.5.3 Description of the experiment

The experiment was designed in the following way. As to avoid previous experience bias a fairly unpopular wiki service was selected: WikiKuchnia[9], which is a wiki-based online cookbook where people contribute recipes in an typical open process according to the wiki method and style.[10] For the purpose of the experiment a new entry was prepared presenting a fictive recipe for a non-existent dish we named 'rogale raciborskie', a fictive variant of meet rolls. Again the dish was fictive as to limit previous experience bias.

Three variants of the wikipage with the recipe were prepared and located on the proxy sever in the domain wikikuchnia.net purchased for the purpose of the study. A mirror of WikiKuchnia was put on the proxy server in WikiKuchnia.net as to avoid disturbance from the genuine WikiKuchnia.org service administrators. Each page variant differ only by the elements necessary for the manipulation. The screenshot of the view of the page in the basic version meant for the control group is presented in the figure 3.1.

The page displayed to the study Group A was manipulated by including a pop-up window with a "WikiKuchnia curiosity of the day" which included a picture of Zuris, big white worms eaten by natives Indians in Peru. This was done under the assumption that such an image would under normal circumstances evoke strong feeling of disgust in people of European culture. The pop-up window included a short information in the bottom lefthand side of the window: "These are Zuris. Indians in Peru eat them alive.". This pop-up was activated in the 8th second after the respondent opened the target wikipage. The pop-up window included an option button to close the window and continue the reading of the wikipage. The pop-up window is presented in figure 3.2.

The third group, the study Group B was directed to a page which was manipulated as well. This time the picture used on the page presenting the rolls was altered.

[9] `http://www.wikikuchnia.org/`

[10] On 18 April 2013, after the experiment the last modification of was timestamped 18:19, 13 January 2012 and the visitor counter was set at 188651

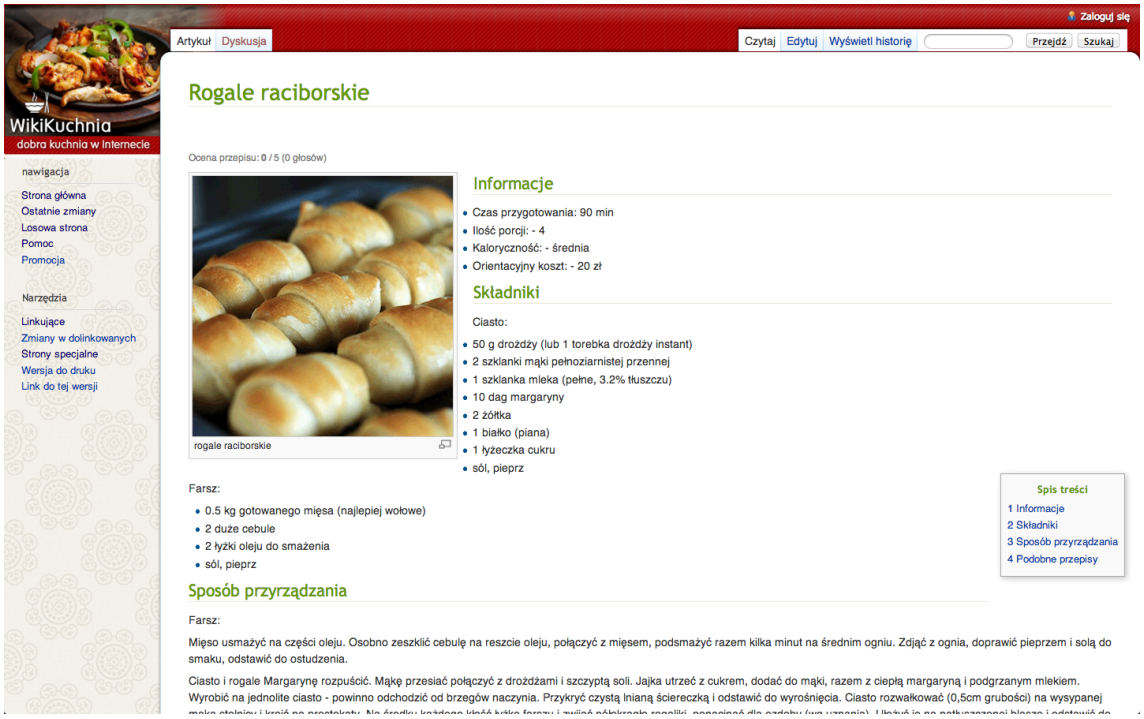

Figure 3.1: Screenshot of the unmodified web page

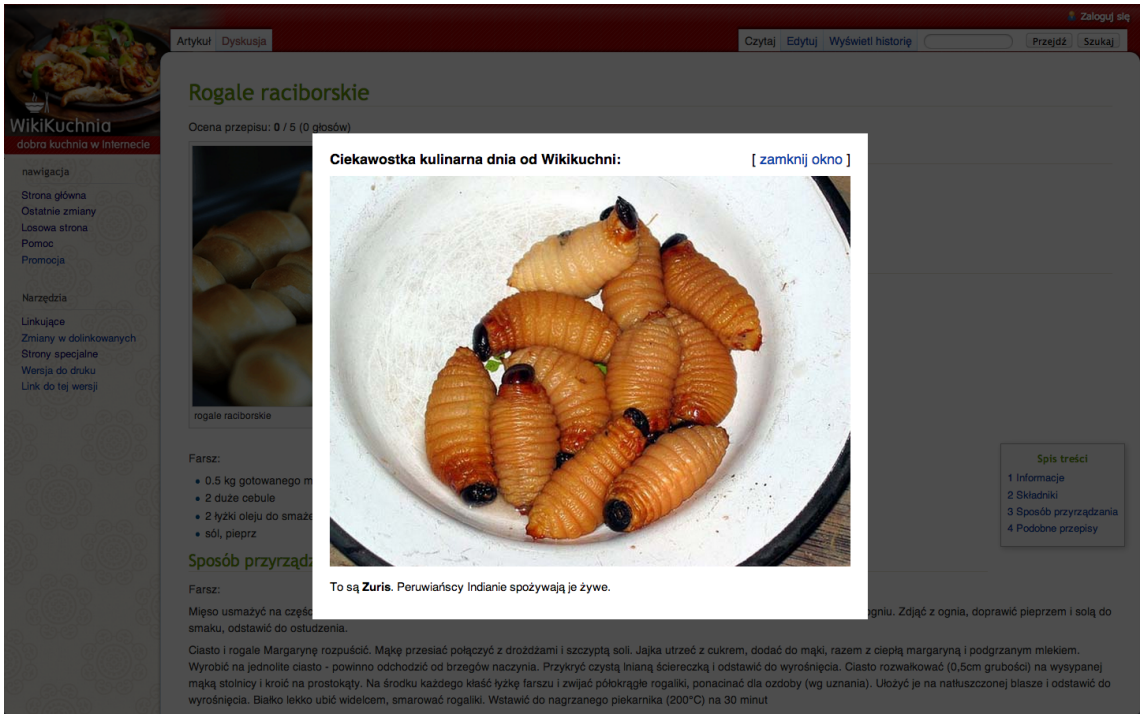

Figure 3.2: Screenshot of the web page displaying a pop-up window

Instead of a regular, static .png file with a photograph of the rolls an animated .gif file was placed which included two layers. One layer included the original picture of the rolls and the second layer the picture of Zuris, the same used in case of the group A. The layers where animated with a delay between frames set at 2200ms for the layer with the rolls and at 17ms for the layer with the worms, looped forever. 17ms delay was chosen as the typical sub-threshold time applied in visual priming (Marcel, 1983; Stone, Valentine, and Davis, 2001). Consequently the subject could only unconsciously appreciate the stimuli evoking the emotion of disgust. It is important to note that in some studies low-latency subliminal priming effect was found to be stronger than in higher supraliminal priming latencies (Fischler and Goodman, 1978).

Importantly, the content of the page in terms of the recipe, structure, layout, etc., was unaltered for all the groups. Each participant was evaluating exactly the same recipe, although participants were informed in the questionnaire instructions that a random wiki entry had been selected from WikiKuchnia, which was the part of the purposeful misinformation. The misinformation was introduced to limit the respondence bias. All participants were informed that the study is about assessing the quality of Polish wiki services. The participants were guided throughout the study and the questionnaire in a way that was supposed to make them believe the qualitative, subjective assessment of WikiKuchnia was the genuine purpose of the study. The third group, the subliminal study group B, was added to the experiment, as to further weaken the respondence bias, as it was anticipated that despite misinformation some participants may still figure out the real goal of the study which would spoil the experiment.

The procedure was as follows. Each participant received a link to the Google Form in one of the three versions corresponding to the particular experimental group. The first page of the questionnaire gathered basic respondent information such as age, gender and a few profile questions irrelevant to the study which were part of the misinformation strategy. Also the respondents were asked about how much they like ingredients which later would appear in the recipe, however this was prior to the moment the respondents could see the wikipage with the recipe for assessment. In the next step, on the next questionnaire page, each participant received the link to the wikipage in one of the three prepared versions. After reading the wikipage with the meat rolls recipe the respondent was asked to come back to the questionnaire and reply the rest of the questions. The questions in the questionnaires where identical for all the groups. Only the links to wikipages where altered. Moreover, the questionnaire for group A included one additional control question at the end of the questionnaire (on the third page) to verify that the pop-up window was correctly displayed on the catered wikipage. The respondents who indicated that they had not seen the pop-up were removed from the data set.

3.5.4 Definition of the variables

The dependent variable under examination was the judgement bias, which we tried to capture by asking participants a set of questions aiming at measuring: (i) qualitative assessment score of wikikuchnia under a set of 5 assessment criteria: clearness, conciseness, structure quality, language quality, visual quality; (ii) perceived tastiness of the meat rolls which we juxtaposed with individual preferences of the ingredients measured before the manipulation; (iii) willingness to order rolls from a restaurant menue (iv) willingness to recommend the dish to a friend, (v) willingness to recommend the WikiKuchnia service to a friend.

As for variable (i) the participants had to evaluate aspects of WikiKuchnia on a 1-4 Likert scale, where: 1 – very poor, 2 – poor, 3 – good, 4 – very good. In case of variable (ii) the participants had to answer a corresponding deliberate question on a 1–5 Likert scale, where: 1 – strongly dislike, 2 – dislike, 3 – neutral, 4 – like, 5 – very much like. Finally, the nominal variables (iii–v) took values [1,2] depending on negative or positive answer to relevant deliberate question corresponding to [No, Yes] respectively.

Details on the variables and the corresponding questions will be provided as the results of the study are presented in the following subsection.

The dependant variable of the study was the presence of the disgust evoking stimulus, i.e. the belonging to the one of the experimental groups: control (1), supraliminal (2) or subliminal (3), which corresponded respectively to (1) the non-presence of disgust evoking stimulus, (2) presence of supraliminally or (3) subliminally projected stimulus evoking the affect of disgust.

3.5.5 Results

284 questionnaires where collected during the study out of which 273 retained, as 11 responses were ignored from the Group B, as in those cases the respondents indicated that no pop-up had been noticed.

The participation structure for the responses retained for the data analysis is presented on figure 3.3.

All statistical tests have been performed with R version 2.15.3. Charts where generated in Google Spreadsheet.

The qualitative assessment of WikiKuchnia

In the study, after the manipulation, the participants were asked to evaluate WikiKuchnia on a set of criteria. The corresponding form question was formulated as follows: “Please assess the quality of the following aspects of WikiKuchnia:” after which 5 aspects followed: clearness, conciseness, structure, language, visual aspects, each to be assessed on a 1–4 Likert scale such that: 1 – very poor, 2 – poor,

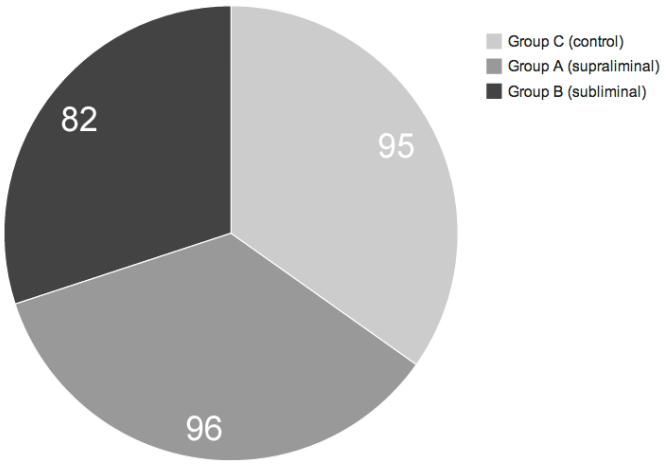

Figure 3.3: No. of participants in the study (retained)

3 – good, 4 – very good. Based on this answers the "average.score" indicator has been defined as the arithmetic mean of the scores for the 5 criteria. Figure 3.4 presents the average score values for all the three experiment groups.

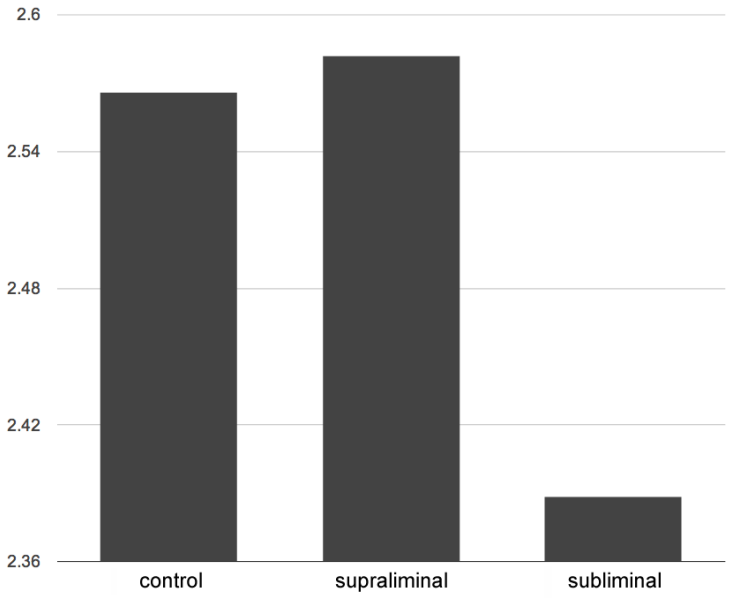

Figure 3.4: WikiKuchnia average assessment score, by group

Visibly there is a difference in average score between the control and subliminal groups, statistical significance of which has been verified by applying between subjects one-way analysis of variance (ANOVA). The analysis revealed a significant effect of only subliminally evoked affect of disgust on qualitative overall assessment of WikiKuchnia (operationalisation: arithmetic mean of assessment scores of all aspects)

at the $p < .001[F(1, 175) = 12.01, p = 0.000666]$. Interestingly this was not the case for the supraliminal group which responded alike the control group. Importantly the Bartlett's test did not show a violation of homogeneity of variances $[\chi^2(1) = 0.4322, p = 0.5109]$.

In order to verify whether the reason for the subliminal group odd responses could lie in the manipulation of the graphical elements of the wikipage, as the animated .gif file with subliminally projected frame containing the disgusting Zuris could potentially impact visual experience of the subjects, the avarage.score indicator was modified by taking out the assessment score for the visual aspects. Consequently a new variable was defined "average.score..without.visual", and statistical tests repeated. As figure 3.5 illustrates the difference between average score values changed only slightly, as means decreased for all groups.

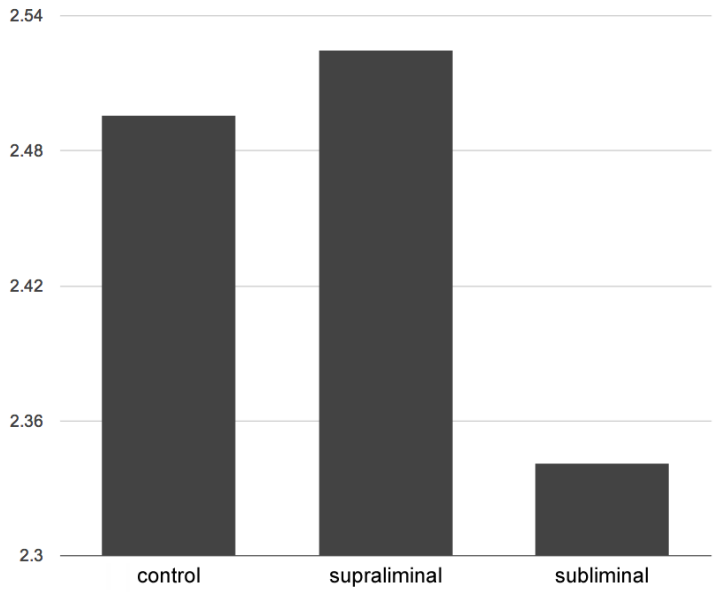

Figure 3.5: WikiKuchnia average assessment score excluding visual aspects, by group

As in previous case Bartlett's test did not show a violation of homogeneity condition $[\chi^2(1) = 0.1338, p = 0.7145]$, and one-way Anova confirmed significant effect of subliminal disgust affect evocation on average WikiKuchnia assessment $[F(1, 175) = 9.278, p = 0.00268]$

The willingness to recommend WikiKuchnia to a friend

In user/customer satisfaction research the question about readiness to recommend a product or service to a friend is the ultimate one. It is assumed to best reflect the true attitude of a user towards the assessed object. For the very same reason this question was pivotal in our study. Under the given experimental design, with the considerably large samples, there should be no difference expected in responses

by the three groups to this question. The analysis of the data from the experiment proves the contrary.

Let us start with presenting the frequencies of yes/no answers for the free groups, summarised in figure 3.6.

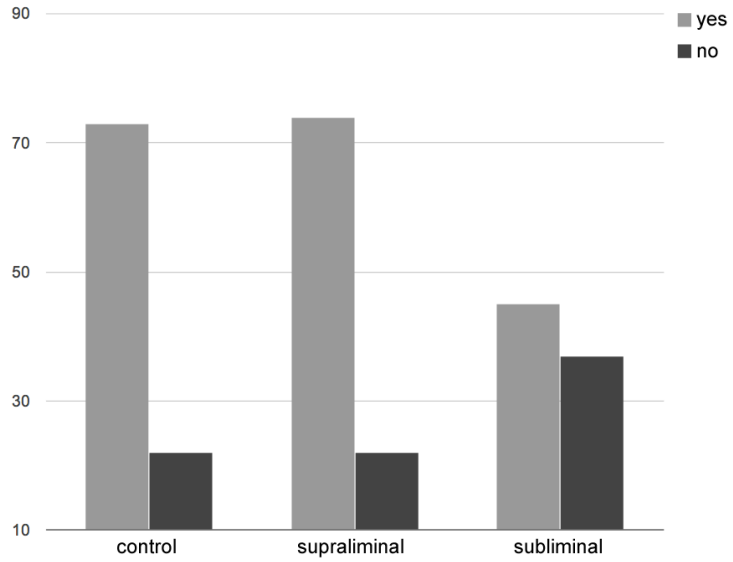

Figure 3.6: "Would you recommend WikiKuchnia to a friend?"

Visibly, the subliminal group stands out again. To determine whether observed frequencies are significantly different between the experimental groups the Pearson's chi-squared test (χ^2) was used. The results of the test confirm that there is a statistically significant dependence between the independent variable (belonging to a particular experimental group) and the willingness to recommend WikiKuchnia to a friend [$\chi^2(2, N = 273) = 13.41, p < 0.01 (p = 0.001224)$]. Again, the dependance only holds between the control and subliminal groups [$\chi^2(1, N = 177) = 9.55, p < 0.01 (p = 0.001995)$], as there was virtually no difference in response frequencies between the control and supraliminal groups.

Subjective assessment of the meat rolls

Another hypothesis of the study was that the affective stimulus used in the experiment would affect the perceived tastiness of the dish presented on the wikipage. This was measured by three indicators corresponding to three questions in the questionnaire: (i) "Suppose you are in a restaurant and you find these meat rolls on the menu among other you know and like, would you order them?"; (ii) "Would you recommend the meet rolls to a friend?"; (iii) "Even if you have not tried such a dish before try to imagine its taste and assess how much you like it". The first questions are yes/no type questions operationalised into a nominal variable taking values [1,2], whereas the

last question involved a Likert scale coded into quantitative variable taking values [1,2,3,4,5] as defined in previous section.

Let us start with presenting the frequencies of the yes/no answers for the free groups. Answers to the question (i) are summarised in figure 3.7.

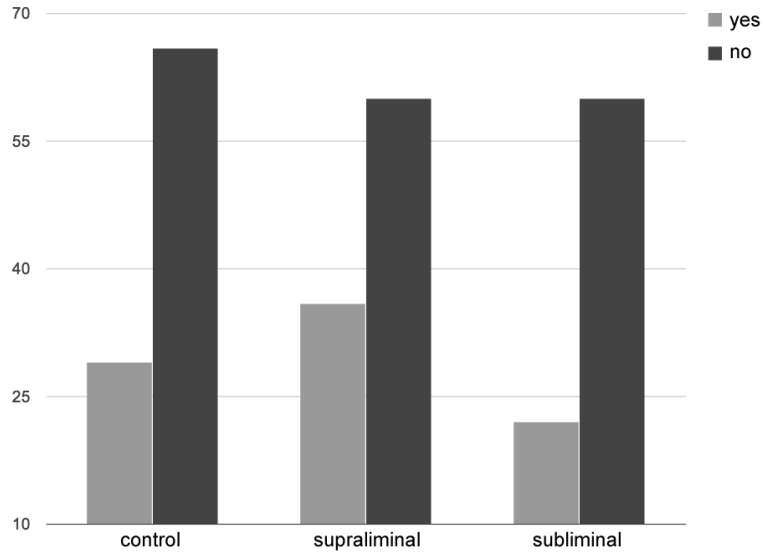

Figure 3.7: "Would you order the meat rolls?"

Although there is a visible deviation in the answers by subliminal group compared to the other two, more precisely subjects from the subliminal group tended to be slightly less willing to order rolls form a restaurant menu, the effect of the presence of the subliminal stimulus is not statistically significant $[\chi^2(2, N = 273) = 2.44, p > 0.05(p = 0.2952)]$.

Now, let us look at the answers to the question (ii) summarised in figure 3.8.

Alike in the case of WikiKuchnia service recommendation subjects in the subliminal group where less willing to recommend the rolls to a friend and this effect is statistically significant $[\chi^2(2, N = 273) = 9.345, p < 0.01(p = 0.009348)]$.

The weaker effect in case of the first question may result from the fact that subjects are more likely to experiment with their own food, and less willing to risk compromising their relationships with a friend by recommending something they potentially dislike or do not know. Importantly the questionnaire included a question on whether the recipe or the dish itself was familiar to subjects, furthermore the recipe was fictional, so the effect cannot be solely explained by subject's avoidance to recommend an unfamiliar dish. As mentioned earlier the question about willingness to recommend an object to a friend is regarded a better indicator for measuring subject experience quality.

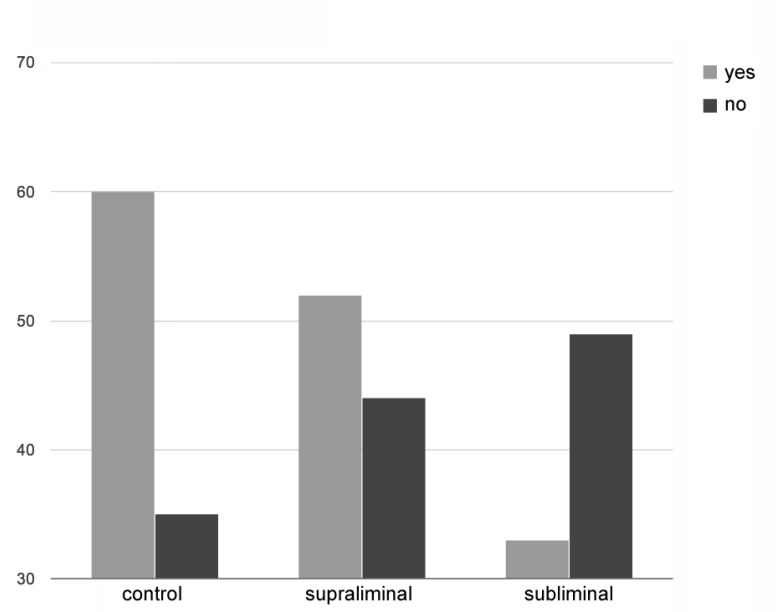

Figure 3.8: "Would you recommend the meat rolls to a friend?"

Finally let us look at the answers provided to the third question measuring the perceived tastiness of the dish presented in the WikiKuchnia entry. Figure 3.9 provides the summary of the mean values by each group.

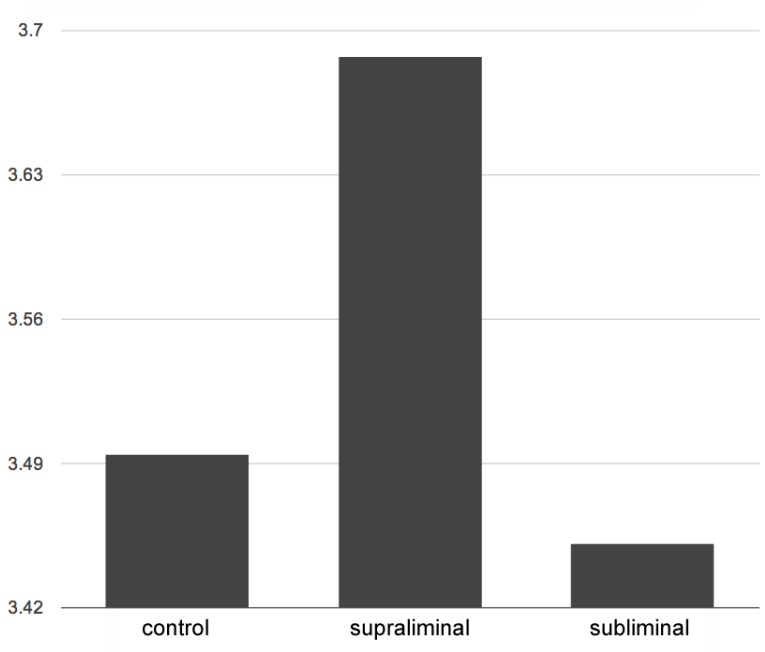

Figure 3.9: "How much do you think you would like the rolls?" (mean by group)

Surprisingly, this time the supraliminal group stands out slightly with higher mean value. However this effect is neither strong nor statistically significant. Bartlett's test did not show a violation of homogeneity of variances ($\chi^2(2) = 0.3526, p = 0.8384$) and the one-way ANOVA test results did not allow to take the effect of disgust on the perceived tastiness of the rolls as statistically significant [$F(1, 271) = 0.068, p = 0.795$]. This also held for non-parametric tests of independence. The contingency table for the considered variable "like.rolls" is presented in table 3.5.5.

Table 3.1: Contingency table for the considered variable "like.rolls"

	control	supraliminal	subliminal
strongly dislike	2	3	4
dislike	10	5	5
neutral	28	17	23
like	49	65	50
very much like	6	6	0

If we apply the Chi-square test to the contingency table we get results which do not allow us to reject the null hypothesis that level of disgust is independent of the perceived tastiness of the dish [$\chi^2(8, N = 273) = 13.4779, p > 0.05(p = 0.09643)$].

As these results are contradictory to our initial assumptions let us consider other elements of the questionnaire related to this point.

In the study, before the manipulation, all participants were asked to declare how much they liked each of the individual ingredients appearing in the recipe that they were supposed to assess later, on a Likert scale 1-5. Our expectations where that: (i) there should be no significant difference between groups in scoring for all the ingredients separately either for the mean of individual scores, (ii) there should be a correlation between the assessment of individual ingredients and the perceived tastiness of the dish for the control group, (iii) there should be a significant difference in perceived tastiness between the control group and the two study groups (supraliminal and subliminal).

We already know that the third expectation turned out incorrect. Now let us look at the other two. Figure 3.10 summarises the average score for all ingredients by group.

The differences in average score between groups are very small, furthermore not statistically significant [F(1,271)= 0, p=0.982]. If we analyse the ingredients one by one and test their dependence on the group with one-way ANOVA we get the results presented in table 3.5.5.

Those results reveal dependence for only 3 out of 13 ingredients, out of which one can be considered as not critical (margarine) in the recipe and for the other two

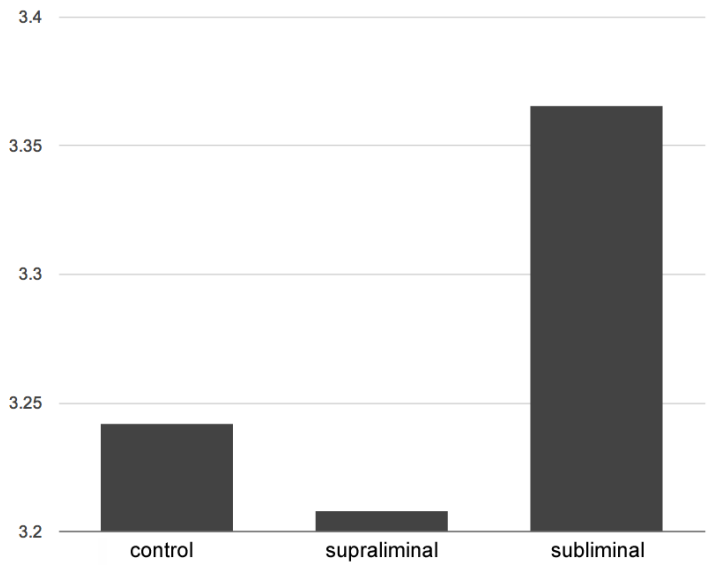

Figure 3.10: Average tastiness score for all ingredients by group

the mean values were higher in the supraliminal and subliminal groups compared to the control group, which is illustrated in figure 3.11.

Consequently, we can say that all the three groups of respondents where homogeneous in terms of how much they liked the ingredients appearing in the recipe.

Interestingly however, there was no correlation between the average assessment of the ingredients before the manipulation and the perceived tastiness of the dish revealed in the questionnaire after WikiKuchnia entry was displayed, for either the control group alone or all the groups put together. Tasting the data for Pearson's product-moment correlation proved that there was no correlation between the average liking of the ingredients and the perceived tastiness of the dish in the control group $[r = 0.23, N = 95, p = 0.02508]$. Neither could it be detected when answers of all the participants where considered $[r = 0.23, N = 273, p = 0.0001203]$. We have received similar result when Kendall's rank correlation tau test was applied [τ=0.2263107, N=95, p=0.005409, τ=0.1724961, N=273, p=0.0003448 respectively for the control group alone, and the three groups together].

The above results suggest that the subjective perceived tastiness of the dish did not depend on how subjects liked the ingredients before the experiment, and it is hard to determine the reasons for why participants liked or disliked the dish at the end, so this indicator is not the best for analysing the deviations in judgements between groups.

Table 3.2: One-way ANOVA test results confirming lack of significant differences between groups in how individuals liked the ingredients before the experimental manipulation

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
pork	1	0.22	0.218	0.367	0.545314	
beef	1	5.25	5.255	8.841	0.003223	**
onion	1	4.18	4.183	7.037	0.008476	**
oil	1	0.57	0.565	0.951	0.330392	
flour	1	0.20	0.198	0.334	0.564060	
yeast	1	0.02	0.017	0.029	0.865620	
milk	1	0.28	0.279	0.470	0.493749	
margarine	1	8.38	8.382	14.102	0.000214	***
yolk	1	0.08	0.080	0.135	0.713170	
egg.white	1	0.05	0.050	0.084	0.772028	
sugar	1	2.23	2.226	3.745	0.054062	
salt	1	0.23	0.233	0.393	0.531496	
peper	1	0.75	0.751	1.264	0.262006	
Residuals	259	153.94	0.594			

3.5.6 Conclusions and further work

One in all the results presented in the previous section suggest that the affect of disgust has statistically significant effect on the rational judgements primarily in situation when subjects are not aware of the affective stimulus. Should the stimulus be consciously appreciated the effect is no longer visible. Furthermore the effect is not strong, as the variations in responses between the control group and the study groups where not high in relative terms. However in case of the key indicators considered in the study, such as the willingness to recommend wikipedia to a friend, the effect was considerable.

The interesting and not expected finding of the study is the lack of effect in case of the supraliminal study group. Members of this group responded very similarly to the control group despite the manipulation. This could be explained by the respondence bias. Once a respondent realises that the displayed pop-up has something to do with the true purpose of the study the experiment is spoiled. We speculate that in such occurrence of events the respondents where particularly cautious and fair when providing their assessment scores, trying to mitigate the effect of disgust they became aware of. This would suggest that in the rational judgment indeed there is an interplay between affective traits and logical, fact-based deliberations, which may to certain degree be consciously controlled.

As for the further work, it would be interesting to repeat a similar study with a within-subject experimental design. This would allow us to verify consistency

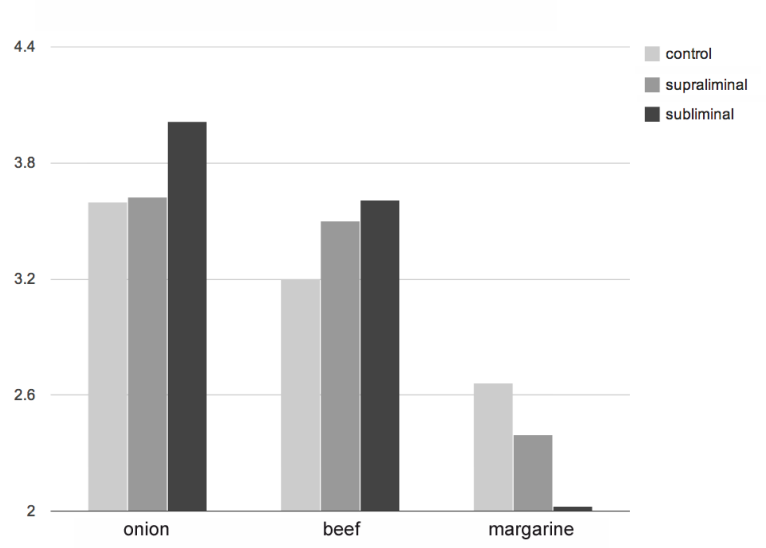

Figure 3.11: "How much do you like these ingredients? (mean by group)"

in answers within subjects, which would allow separate better participants' past experience bias. Such design however would call for a more controlled experiment environment and thus require more resources for equally high participation numbers.

3.6 Conclusions

Do we happen to leave in increasingly romantic times? Shall we now turn to the classics of romantic literature and thought for inspiration? Or perhaps it is all just a puzzle, and the key to this puzzle lies in the proper framing of the problem? Evidently the dualist tradition makes the boundaries between mind and body, heart and reason, rational and emotional clear and sharp. The mind-body problem has been approached by contemporary philosophers by objecting dualism, noticing that the same reality can simply be described from different narratives: mental and physical that are not mutually exclusive. The dualist tradition in considering the rational as separate from the emotional is strong as well, which is mainly driven by classical and neo-classical economics. Likely, the solution again can be found in questioning the traditional dichotomy between rationality and emotionality, dissolving the positivist-romantic problem in the integral spirit.

Let us conclude this chapter by highlighting the following characteristics of subjective affective mental states: (i) affective mental states have biological representation; (ii) affective states are subjective states but can be compared inter-subjectively; (iii) the problem with affective states representation begins with the limitation of natural language which prompts for the search of non-linguistic forms of intentionality

Elaborating on the last point let us note that traditionally what has been described as rational linguistically pertained to the realm of science and philosophy the emotional pertained to the realm of arts. This is exactly because these two realms are governed by different narratives, the former by the intra-subjective and objective reality and the former by primarily subjective reality. Speech acts are forms of institutional intentional states therefore linguistic components have corresponding parts of objective reality that they represent. Everyone knows what a "wheel" is because we can point at an object that you and me see and say "This is a wheel". We cannot do this with affective states because the they are subjective and likely private, so it is hard to define the "direction of fit".

Let us close the discussion in this chapter with two quotes from Damasio pointing at two very important properties of of conscious mind states:

> "[Conscious states of mind] contain an obligate aspect of feeling - they feel like something to us."

> "Of necessity, conscious states of mind handle knowledge based on different sensory material - bodily, visual, auditory, and so forth - and manifest varied qualitative properties for the different sensory streams. Conscious states of mind are *felt*" (Damasio, 2010)

Chapter 4

A model for experience representation in information systems

The aim of this chapter is threefold. Firstly (i), it will introduce the conceptual framework for unified representation of affect and knowledge in informations systems. This will allow (ii) to introduce the formal definition of experience and (iii) propose experience representation framework for application in information systems. Meanwhile the said framework will build on the mainstream knowledge representation approaches without entering into a discussion on their validity nor efficiency, the complementing part of the model corresponding to affect will be studied in more detail. To be more specific (iv) a review of emotional and affective models used in information science with the objective to identify their main weaknesses and limitations will be carried out. Finally, we will be able to conclude with proposing a general purpose theory of experience applicable to representation, emulation and processing of experiential phenomena in information systems.

4.1 From knowledge representation to experience representation

Although knowledge representation is an established discipline in artificial intelligence research and information science, representation of experience is a new concept. Queries on "experience representation" in mainstream scientific resource search engines such as Google Scholar return few, mostly unrelated, results, which suggests we are opening a new chapter in this respect.

Obviously, there have been efforts aimed at representing emotions in computer systems, under a field of affective computing/reasoning that has emerged recently. Affective computing embraces the studies on computational models of human emotional processes that mostly serve the purpose of modelling human emotional behaviour in

computer and robotic systems. Affective computing is a relatively young discipline involving mostly computer scientists and cognitive psychologists that has grown out of cognitivist tradition. The cognitivist legacy of the field has had a profound impact on the shape of the mainstream theories, models and affective representation frameworks. It manifests in the popularity of the so-called appraisal-based models of emotions worked out over 80s and 90s of the XX century by cognitive psychologists, with most prominent names including Lazarus (Lazarus, 1991), Scherer (Scherer, Schorr, and Johnstone, 2001) and Frijda (Frijda, 1986).

Before providing a review of main theories and models of emotions underpinning the conceptual frameworks of computational models of emotions, which will be included in the following section, let us establish a link between experience, knowledge and emotion representation.

The said link is quite straightforward and is a direct consequence of the way we understand and define conscious experience. In order to have a sound representation framework of experience, which we define as affectively coloured knowledge, one needs to have two basic elements to start with: (i) a sound knowledge representation framework; (ii) a sound affect representation framework. In order to have the latter one needs: (iia) a sound theory and model of affect, (iib) a proper formal (computational) model which can implement the theoretical model of affect. The (i) and (ii) could, and ideally should, be addressed by one unified framework for dealing with experience as a unified phenomena.

Importantly as for (i) knowledge representation that covers the representation of intentional content of mind in IS there are already proven models and methods provided by knowledge engineering field. We will rely here and in the future work on these proven models. However, There is a more apparent lack of (ii) affect representation frameworks as well as gap-bridging between (i) and (ii). So the natural starting point for us in improving the situation is to challenge (ii) in fist place, therefore (iia) and (iib).

We find that neither a sound, usable theory of emotion (iib) nor a formal framework to capture affective phenomena in IS (iib) are in place currently. To us, the lack of (iia) seems to be at least partly due to a particular legacy of cognitivist thinking that has dominated the AI field. AI researchers have relied on the "armchair" type accounts of emotion provided by psychologists and cognitive scientists without much critical reflection. Only recently, and surprisingly slowly, new approaches based on accumulated scientific evidence form brain science has started making the wave and has been attracting more and more attention from information and computer scientists, but no substantially new paradigms has been proposed so far.

The lack of (iib) is too large extent a consequence of refutable (iia), but also stems from the fact that mainstream KR formal systems are intrinsically objective, episteme-fixated and language dependant. Meanwhile it is difficult to disagree with Panksepp who claims:

> "Affects are not encoded as information. They are diffuse global states generated by deep subcortical brain structures, interacting with primitive viscerosmatic body (core self) representations that remain poorly mapped."(Panksepp, 2008)

The natural tendency in the field of information and computer science in approaching affective phenomena is to apply proven knowledge engineering mode of thinking, methods and techniques to affect representation, despite the fact that knowledge and affect are distinctively different phenomena, which is a pitfall in which we also prone to a certain degree, but we realize it and have taken it as our main direction for improvement and further work. Possibly, a fundamentally new formal system or even a new type of formalism may be needed to facilitate a development of a satisfactory (iib).

The overview of theoretical approaches to emotion modelling and representation in information system, to be covered in the following subsection, shall provide justification of the above formulated diagnosis. Importantly, we will focus in this review primarily on identifying weaknesses of the emotional and affective theoretical models providing basis for computational models of affective phenomena because we strongly believe that the former are the main reason for inefficiencies of formal, computational models. Formal models always result in yet another degree of simplification of the natural phenomena. If the formal models are based on refutable theories and theoretical models of emotion and affect they are a priori bound to failure, this is why we emphasise the need for coming up with an improved theoretical model before getting to implementation details. Applicability of the theoretical framework proposed by us later in this chapter will be covered in the next one.

4.2 Limitations of mainstream affect and emotion models

The recent review of contemporary computational models of emotion by Marsella et al. (Marsella, Gratch, and Petta, 2010) indicates that there are, at least in principle, three major theoretical traditions underlying the formal models: (i) appraisal theory, (ii) dimension theories, which in practice has been often applied to formalisms relying also on appraisal theory, (iii) anatomical and (iv) rational. Unquestionably, the most popular and influential among the authors of computational models of emotion, is the first theory (i). Figure 4.1 provides a detailed mapping of all models considered by the cited review onto the affect/emotion theories that provide theoretical and philosophical background for formal representation of affective and emotional processes in information and computer systems. The dominant position of appraisal theory tradition is apparent. This tradition has been shaped by four theories of emotion: (i) Appraisal theory by Lazarus (Lazarus, 1990; Lazarus, 2006), (ii) sequential checking

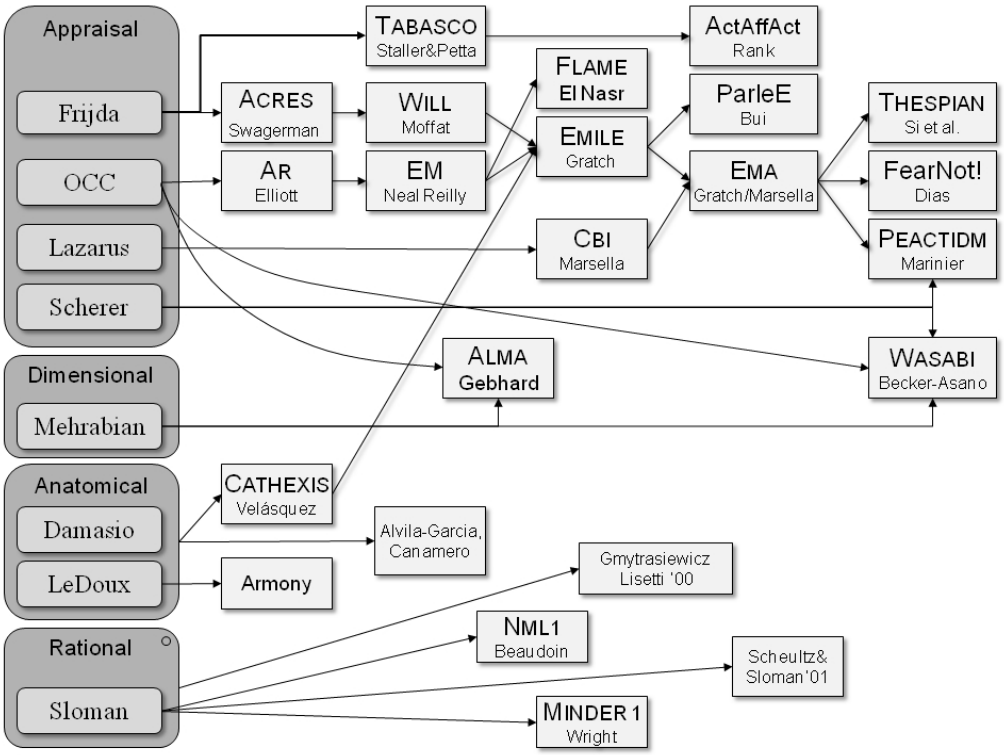

Figure 4.1: Overview of contemporary computational models of emotion by Marsella et al. (Marsella, Gratch, and Petta, 2010)

theory by Scherer (Scherer, Schorr, and Johnstone, 2001), (iii) action tendency by Frijda, (iv) OCC model by Ortony, Clore and Collins. All four have two important things in common: firstly, they treat emotion process as a cognitive process that starts with appraisal of the triggering stimulus producing, in course of conscious reasoning, behavioural or cognitive responses as process outputs, secondly, they confuse emotion and feelings of emotions.

According to the evidence presented throughout Chapter 3 these theories are flawed. We will briefly present Lazarus's version of appraisal theory as an exemplification of this family of theories as to highlight the main flaws, later we will concentrate on OCC model chosen as an example of appraisal theory driven emotion taxonomy and undertake its critical review. Before that, let us mention that many propositions that build up these theories are uncontroversial and plausible, such as that emotions are adaptation mechanisms produced by evolution, and that their predominant role is to help organism to ensure survival and well-being. Without a prejudice to the informative and useful insights provided by these theories the below critical presentation will concentrate on more controversial aspect of appraisal theories of emotion.

The account of emotion developed by Lazarus is the most widely cited among all appraisal approaches. Like all theories in the family it proposes that:

> "Each emotion expresses a person's appraisal of a person-environment relationship involving a particular kind of harm or benefit."

(Lazarus, 1990, p. 611) According to Lazarus the appraisal involves *interaction* between antecedent motivational and belief variables on the one hand and environmental demands, constraints and resources on the other. The appraisal leads to generation of embodied *action tendencies* (en element strongly emphasized also in Frijda's account) that correspond to the construed harm/benefit relationships. Importantly Lazarus states:

> "Above all, the emotional response is not a reaction to a stimulus, but to an organism(person) – environment *relationship*."(Lazarus, 1990, p. 614)

This statement is critical as it shows that according to the discussed theory only an interpreted stimulus can trigger an emotional state. The interpretation that is done in the appraisal phase is a cognitive process that involves conscious thought processing. This contradicts findings by Ledoux (LeDoux, 2000) on emotional fear responses discussed in earlier chapters as well as the phenomena of inbred fear reaction to certain objects such as predators like cats in case of laboratory rodents (Panksepp, 2005), or human innate fear reactions to snakes (Öhman, Flykt, and Esteves, 2001). This also is opposed by an account of Zajonc (Zajonc, 1980; Zajonc, 1984), who suggested that occasionally emotion is independent of cognition.

The basic proposition of the appraisal theory that we find fundamentally wrong is that emotional states are only triggered by cognitive information processing. Lazarus proposes that:

> "[Representations of one's circumstances], which reflect knowledge and beliefs about what is happening, are relevant to emotion because they are the data that the person evaluates with respect to their adaptational significance. These knowledge-centred representations, or *situational construals*, however, do not directly produce emotions. Instead, it is how these representations are appraised with respect to their significance for personal well-being – the second type of cognition – that directly determines the emotional state."(Lazarus, 1990, p. 616)

The above provides an account of an emotional reaction as a complex cognitive process requiring involvement of high level cognitional capacities of relational abstract thinking. Under such and account an emotion arises only when an agent perceives an event in which certain environmental configuration and personality traits come coincide to impact its subjective well-being. The process is illustrated by figure 4.2. Let us comment on one selected instance of the process illustrated thereof, namely the creation of "affect", i.e. the subjective experience of the natural agent experiencing an emotion. Little is said about this component and the consequences of it are not clear. There is no place in the model for association of affect with knowledge and purposeful behaviour, affect is presented as a phenomenal irrelevant by-product of the process which have to be "coped with". This is exactly the moment when the theory ignores the importance of feelings of emotions and leaves the relationship between emotion, emotional response and feelings of emotions unclear. It is astonishing as under such an account there is not much justification nor the role for affect to exist. Why would evolution create affective dimension of experience if there is no practical role for it. We propose that this is a critical moment in which the appraisal theories miss the point of long-term influence of emotion on knowledge and future cognitive states. The relationship between feelings of emotions and the entire process as well as each particular instance is memorized and must make a difference in future situations, pretty much like the somatic marker hypothesis by Damasio suggest but differently framed for the purpose of knowledge representation research. We will propose in later parts to address this by introducing a concept of affectively coloured knowledge that will serve as a vehicle for experiential continuity and consistency over time.

Apart from the above flaws the appraisal approach would provide otherwise plausible account of emotional machinery, as it could be patched with introducing the concept if unconscious processing, avoiding thus the troubles with the inefficient and resource intensive conscious cognitive processing, riding on the valid ultimate goal of emotional processes to perform adaptive behaviour in ensuring survival and

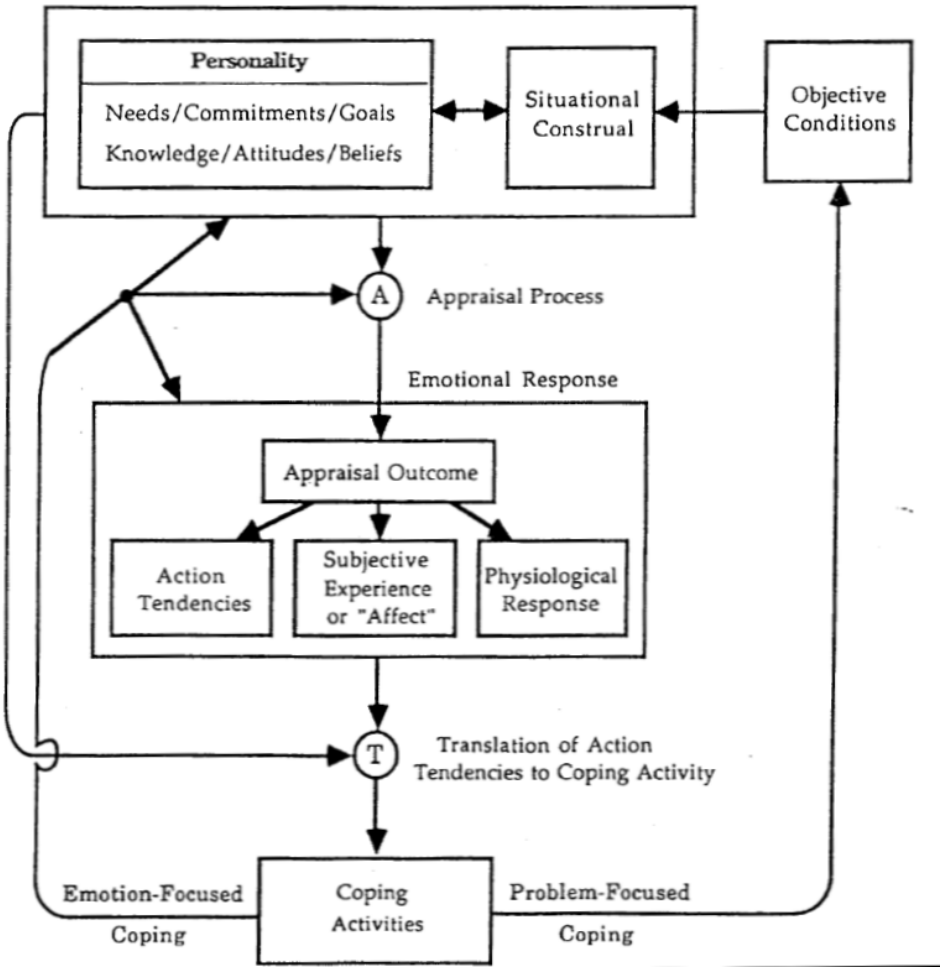

Figure 4.2: The model of the cognitive – motivational – emotive system by Lazarus (Lazarus, 1990)

well-being of an organism. This the case of a modern variant of appraisal account of emotions as formulated by Stein et al.:

> "From the very beginning of life, experiencing and expressing emotion are goal based, expressive, and action-oriented. Experiencing emotion involves continual monitoring of personally relevant goals, and involves a constant appraisal of the value and worth of events, people, objects, activities, ideas, internal states, and anything else that impinges on the psychological and physical well-being of the person involved. The monitoring of goals requires both unconscious and conscious processing, and the evocation of emotion is the result of both types of processing. One of the main questions surrounding the description of the thinking that accompanies emotion is when and under what conditions emotional processing results in conscious awareness, as opposed to remaining an unconscious process." (Stein, Hernandez, and Trabasso, 2008, p. 576)

The above theory of emotion as any other appraisal based approach has been strongly influenced by Miller's classic account of the importance of planing and goal achievement for human behaviour (Miller, Galanter, and Pribram, 1960). For Stein *goal monitoring* is from where emotional excitation originates. This is in line with the commonly accepted claim that emotion primarily serve evolutionary fitness so they must help an organism to guide its behaviour in a way that this principal goal such as survival are attained. The trouble with such an account is if the nature of all types of goals, such distinctly different as thirst satisfaction, survival and going to cinema tonight, can be *monitored* and *appraised* with the same sort of mechanisms. Despite the theoretical elegance of Stein's account it is hard to ignore the important evidence from neuroscience that suggest that emotions, some at least, are rather immediate responses to sensory inputs that directly trigger emotional state without entering neither consciousness nor areas in the brain that are capable of cognitive appraisal. Basic emotions are governed simply by lower structures in the brain that are not cognition capable, and the emotional appraisal mechanism seems to be much more primitive than appraisal account would suggest.

Lazarus however, in considering the appraisal mechanism, surprisingly distinguishes emotions from other "entities that serve adaptive purposes" in which he includes (i) "reflexes" such as startle or eye blink, and (ii) "physiological drives" such as thirst or hunger. Distinguishing reflexes from emotions would not be surprising if Lazarus would not suggest in one place that reflexes may also be referred to as "instincts" and assign them to simple organisms "that can afford to interact with their environments in highly stereotyped ways". Unfortunately Lazarus does not specify how "simple" organism he means, whereas these are primitive single cell organism or more complex organisms such as insects or reverberates. The word "instinct" suggest that at least primitive species of reverberates like amphibians are

included. If it is so, the manoeuvre with introducing reflexes as separate entities allows him to avoid inconvenient at the time comparisons of emotional states across species, in particular between human and animal emotions. It serves as a security valve for the theory whenever a visibly similar emotional state is observed in a human being and in an organism that have not sufficient *cognitive capacities* to accommodate the emotion theory proposed by Lazarus, which is quite demanding as far as brain capacity requirements are concerned. In other word a simple organism could not simply experience fear as it would not have enough cognitive capacities to produce fear response, so in this case Lazarus would suggest that the animal is not experiencing fear but launches a reflex. We know from evolutionary studies on central nervous system that this is not correct. Both human and animal emotions, at least the so called basic emotions, are handled by evolutionary older, more primitive in evolutionary sense, parts of the brain.

In general the evolutionary perspective on emotion as outlined by Lazarus is not plausible, as he suggest that in course of evolution "judgement took over from innate reflexes, and emotions-drawing upon both motives and thought-have become the key adaptive process intervening between environmental challenges and actions". Further on Lazarus states: "(...) more complicated species have to stake their security on the capacity to evaluate the significance of what is happening."(Lazarus, 1990)

This contradicts the evolutionary brain scientific account of emotion (compare (Panksepp, 1998)) which is supported by solid experimental and neurological evidence. Somewhat similar is the case of the aforementioned *physiological drives*. Lazarus treats them as entities of adaptive purposes that are parallel to emotions and entail direct behavioural consequences. According to Lazarus account the drive leads directly to behaviour which is adjusted to situation due to the learning capabilities of organisms. So drives constitute sufficient motivational condition for prompting organism action aimed at satisfying them, and the mechanism for selecting the relevant action are dependant on learning and in some higher drives typical of humans such as "to explore, achieve, and gain mastery over the environment as well as to maintain contact and form social bonds with others" depends on "a powerful and abstract intelligence". The omission of emotions in this chain, rather treating them as separate entities is astonishing. From the brain science perspective drives, or what corresponds to the entity described by Lazarus, merely serve as trigger to emotions that motivate organism to seek for resources. It has been known for some time already that a dedicated "seeking" emotional system plays a critical role in this process and that there is a dedicated general purpose neuronal system in a mammalian brain responsible for orchestrating this affective state (Panksepp, 1998). Only because this emotion is triggered by thirst or hunger, which indeed stems from the bodily changes particular to a given homoeostatic need, the animal finds energy and motivation to get the resources, water and food it needs. The most convincing argument for such explanation, contrary to what Lazarus proposes, is the clinical case

of patients suffering from severe depression, in which despite the presence of evident homoeostatic needs of hunger and thirst reflected in the parameters of the body such as low level of sugar and density of blood, these patients do not find enough motivation to get up to take food or drink. Noteworthy, it has been demonstrated that the activity of general-purpose seeking emotional system in the brain responsible for "go get resources" action motivation is chronically low in depressed patients (Panksepp, 2005).

To sum-up the above part of the discussion on the appraisal approach let us conclude that despite theoretical elegance and completeness of this account and its alignment with the evolutionary perspective on behavioural adaptation to environmental conditions guaranteeing long-term evolutionary fitness of the organism this account of emotion is incompatible with a bulk of neurological evidence suggesting that affective states are triggered a the level of brain structures that are not capable of complex cognitive processing which *goal monitoring* would require. It further ignores the fact that affective states that start to develop before conscious cognitive processing influence, so *affect*, the later stage modulation in higher brain structures. This suggests that affective states a priori *colour* beliefs which form the body of knowledge used as a basis for thinking and higher level processing. Despite that the appraisal account remains still the dominant theory of emotion adapted by information and artificial intelligence scientists. This is also because the appraisal account gave birth to a number of emotion models that served well the purpose of emotion representation in computer systems, of which a most influential example is the OCC model presented in the next paragraph.

An important landmark in modelling emotion in artificial systems was the appearance of the seminal, and still widely used in AI field, work by Ortony et. al. (Ortony, Clore, and Collins, 1988), in which a structural framework for dealing with emotions as cognitive phenomena was proposed, labelled OCC after the first letters of the authors' surnames. The unquestionably valuable contribution by Ortony and colleagues was to bring the importance of emotion to human cognition to the attention of AI field widely dominated by computer functionalism at the time, where the theory of human mind as information processing machine was in the mainstream. The introduction to the book opens with the following statement:

> "Emotion is one of the most central and pervasive aspects of human experience."(Ortony, Clore, and Collins, 1988, p. 3)

This work however took one step further which was critical for its popularity. It proposed a ready-to-use theoretical framework allowing to treat the emotion as part of the cognitive machinery of human mind, which has set the thinking pattern on emotion in AI for years to come. Importantly however the cognitivist account of emotion as proposed by Ortony et. all has proven fundamentally wrong. The below quote summarises this account.

> "The emotions are very real and very intense, but they still *issue from* cognitive interpretations imposed on external reality, rather than directly form reality itself. It is in this sense that we claim that there is an essential and profound cognitive basis for emotions." (Ortony, Clore, and Collins, 1988, p. 4)

As mentioned earlier contemporary brain science proves the contrary the basis of emotion is not cognitive but neurologically, unconsciously hard-wired. Emotions are cognitively triggered and may be cognitively modulated but are *not* cognitively determined, they are neurologically determined, automatic programs of bodily responses to stimuli which may in some cases operate entirely without involvement of the higher level cognitive machinery of human brain.

The characterisation and structuralization of affective dimension of human experience proposed by Ortony at al. was driven by the cognitivist paradigm which may be the roots of the problem and the reason why eventually it fails to capture the true nature of feelings and emotions of a natural agent. Under OCC affect is understood as "evaluative reactions to situations as good or bad" and emotions, treated as one kind of "affective conditions" as "valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is construed." In other words the reactions affected by stimuli depend entirely on the stimuli and the evaluation by the natural agent which reminds a deliberation process and involves complex reasoning including that of distinguishing which is good and bad. The complexity and dependence on practical reasoning typical of higher cognitive brain functions is well illustrated by the below excerpt:

> "The qualitative nature of the affective reaction depends in the first instance on what aspect of a situation is evaluated: an event, its agent, or an object. Depending on which of these is the focus of attention, the primary affective reactions include being pleased or displeased, approving or disapproving, and liking or disliking. In particular, the reaction of being pleased or displeased reflects one's *perception of the consequences of events and desirable or undesirable*. Desirability is computed on the basis of the implications an event appears to have for one's goals. The reaction of approving or disapproving reflects one's perception of an agent's action as praiseworthy or blameworthy. Praiseworthiness is computed on the basis of the standards, principles, or values implicated by the action. finally, the affective reaction of liking or disliking reflects one's perception of objects (including persons, things, ideas, experiences, etc.) as appealing or unappealing with respect to one's attitudes towards them. (Ortony, Clore, and Collins, 1988, p. 191)

We know that emotional responses simply do not work this way and have a direct and immediate impact on natural agent behaviour. One sentence from the above

quote is particularly worth commenting: "Desirability is computed on the basis of the implications an event appears to have for one's goals". This pretends to reduce emotions to symptoms of certain cognitive states, which only indirectly influence behaviour. What would such a emotional intermediary be for at all? What would be the evolutionary justification for such an implementation in human body, why would we need emotions in general is that case? Furthermore, there seems to be a recurrence deadlock in the above account, because it is not clarified how the priority of goals is set. If practically reasoned goals are the ultimate driver what decides about that one person has a goal to listen to Mozart and another to Chopin instead? If we accept this stance we end up in the behavioural nonsensical account of a human being as a goal pursuing zombie, which is wrong.

The most significant contribution by Ortony et. al. was their cognitivist emotion typology. Having rejected the basic two-dimensional description of emotions along the typical *valence* and *arrousal* axes for being too "uninformative" and "unsurprising", they proposed that the overall emotional system can be structured into groups or families of emotion types that share the same eliciting conditions. This led them to the division of emotions into three main branches representing respectively the valenced reactions to *events*, *actions* and *objects*. Further on they distinguished six major groups of emotion types taking into account the temporal dimension, relationship to self and other agents as well as objects of external world, and desireability. Consequently thus defined six emotion types embraced 20 emotions forming 10 pairs: (i-ii) love–hate, (iii-iv) admiration–reproach, (v-vi) pride–shame, (vii-viii) joy–distress, (ix-x) gloating–pity, (xi-xii) happy-for–resentment, (xiii-xiv) satisfaction–fears-confirmed, (xv-xvi) relief–disappointment, (xvii-xviii) gratification–remorse, (xix-xx) gratitude–anger. Eventually, an individual structure for each of the 20 particular emotion has been proposed.

The criticism of the OCC model could fill up a separate volume, for which we do not have enough space here. Let us just state that the model contradicts many findings about the nature of emotions and relies on false assumptions such as for instance that emotions depend on whether they are triggered by event, object or agent. It it was the case the fear response to falling down, seeing a horse running on you or viewing a picture of a scared face would be different, and we know it has the same foundations.

The OCC framework is flawed upon yet another important element, it is ambiguous about distinguishing feelings of emotions and emotions, or rather it dubs the general confusion on this issue common at the time. Although the OCC model could be potentially adapted to the needs of representing emotional reactions (execution of the "emotional programmes") it entirely ignores the qualitative, subjective feelings that are invoked by emotions. The proper representation of interdependence between emotions on the one hand and feelings of emotions on the other remains a challenge for AI.

As the OCC model was created with the purpose of AI applications the model is "elegant" enough to be computationally tractable. For the reason of being tractable and highly popular as it has been implemented in vast number of artificial agent frameworks it is often accepted by KR theorists unthinkingly. This becomes evident after the review of many examples of agent logics looked into in detail in the following chapter. The common justification for adopting OCC model of emotion is exactly as follows: as the model has been created with the purpose of IA application, is commonly used by AI researchers and is easy to implement in computing machines we choose to formalize it also in our work. This approach is highly misguided, as the foundations of the OCC model are wrong. This is why we believe there can be no breakthrough in affective computing and affective information systems until a computationally tractable yet *valid* theory of emotion is proposed and made popular.

Needless to say, a flawed account of affect lying at the foundations of OCC model does not have to necessarily disqualify the contribution by Ortony et al. at large as their structuring approach and parts of the proposed framework could serve as a starting point for improved structuring efforts. As the limits of the framework are evidently rooted in the misguided foundations these suggests the most promising directions for improvement. It is also important to recognize that the work by Ortony and colleagues vastly contributed to the convergence of disciplines working by then on emotions separately: psychology, clinical psychology, cognitive science, information and computer science and economics. It was and important landmark in the emergence of what is now called the affective computing.

Affective computing is a discipline which studies the representation of emotions in information and computer systems, that has emerged about the time of the publication by Picard (Picard, 1997), which starts with the following manifest.

> In this book I will lay a foundation and construct a framework for what I call "affective computing", computing that relates to, arises from, or deliberately influences emotions.(...) Affective computing includes implementing emotions, and therefore can aid the development and testing a new and old emotion theories.

Although it is relatively young discipline we have witnessed a wave of publications in the area touching upon many application areas of which most common are: computer games and animated video (believable human agents), human-computer interaction, human behaviour simulation (including economic behaviour), software engineering (intelligent emotional agents), and validation and testing of emotion theories.

The second group of theoretical approaches to affect modelling includes the so called dimensional theories (Marsella, Gratch, and Petta, 2010). The basic idea behind these accounts is that emotions are not discrete, instead it is believed that affective states are like vectors in multidimensional space that embrace the entire domain of possible affective states. The exact number and labels of the dimensions

in the space vary from author to author, yet the most accepted and uncontroversial is the two-dimensional space where one dimension is affective *arousal* and the second *valence*. One of the first and most influential dimensional model was proposed by Mehrabian and Russell which included three dimensions: (i) pleasure-displeasure, (ii) arousal-disarousal, and (iii) dominance-submissiveness, referred to in the literature as PAD model. (Mehrabian and Russell, 1974)

Importantly the Mehrabian–Russell model (M–R) assumed that affects orchestrate human-environment relationship, so that stimuli coming from the environment impact emotional state of an agent which directly elicits given behavioural response. The mapping of environmental events onto behavioural responses mediated by emotions has still been however under a visible influence of the behaviourist way of thinking. Although a body of neurological evidence has been recalled to support the theory the proposed framework severed rather the descriptive purposes of behavioural responses than understanding the nature and neurological roots of emotional behaviour.

The PAD model has been extensively explored in computational models of affect, although it has evolved into many different variations. Russell and Pratt (Russell and Pratt, 1980) later proposed that the dominance dimension, which stands for a degree of power or control, could be omitted in the basic variation of the M–R model, as arousal and valence dimensions are sufficient to cover all possible responses to all types of situations. At the same time Russell proposed an extended framework that embraced with a circular spacial model eight affective concepts: pleasure (0°), excitement(45°), arousal(90°), distress(135°), displeasure(180°), depression(225°), sleeplessness(270°) and relaxation(315°)(Russell, 1980). Recently however Russell has been insisting on the fundamental role of the basic two dimensions: valence and arousal proposing his concept of *core affect* understood as central "states experienced as simply feeling good or bad, energized or enervated", which "influence reflexes, perception, cognition, and behavior and are influenced by many causes internal and external, but people have no direct access to these causal connections." (Russell, 2003, p. 145), see also (Russell and Norvig, 2009).

Another important particularity of the dimensional account is that it highlights the primacy of emotion over cognition in agent–environment instant relationship orchestration. Unlike appraisal theorists Russell believed that the sequence of affective response is as follows: external stimulus occurs, the affective quality of this occurrence influences the core affect, the variation in the core affect is attributed to the stimulating object and only then the cognitive appraisal of the object is carried out in terms of goal relevance. (Russell, 2003)

Regardless of the exact final shape of the affective dimensional space the pivotal element of this account is the questioning that emotional programmes or affective states are *natural kinds*, i.e. entities that have constant properties independent of us humans, but rather resulting from the nature of things. Barrett for instance provides and overview of evidence against the view that there are kinds of emotion

with boundaries that are carved in nature (Barrett, 2006). So there is no universal meaning of the term sadness or joy as these have not objective, natural properties. Such an account contradicts however the anatomic approaches that are inspired by brain research and try to mimic brain processes when modelling emotions. The anatomic account, though we prefer to call it *neuroscientific* account of affect will be briefed in below paragraphs

Let us start with a quote from Panksepp, who's account of affective phenomena appeals to us as most convincing:

> "The basic emotional affects are primary brian/mind processes, similar to seeing a color. One can use a word, like "red", as a label for a color, but this word does not explain the experience of seeing red. (...) In order to explain seeing red, one must discover the neurophysiological and neurochemical causes of visual experience."

Neurological account of emotion is exactly about finding the neurological and neurochemical correlates of all variety of affective states by studying human and animal models with all sorts of techniques that brain science has at hand: brain imaging, electrical and chemical neurostimulation combined with behavioural experiments. As such is seems the most scientific way of studying emotions as it operates with objective categories. Indeed, what Panksepp postulates above is that this is the only way to study emotion scientifically.

The basic idea behind neurological account of affect is that affective states are results of neurological processes governed by brain structures that can be mapped and explained. These processes are hard-wired, if not subject to brain plasticity, and can be separated according to variety of functions they fulfil in life management of an organism. It also, at least implicitly, assumes that as equalities between these processes in brains of different individuals and sometimes across species can be identified the subjective quality of experiences produced by these hard-wired affects are likely to be equal. This assumption has two important implications: (i) studying of brain processes behind affective states can provide us with meaningful clues on the subjective properties of these states, (ii) affective states are intersubjectively comparable, as their neuronal correlates can be objectively compared.

Another important implication can be noted from the earlier quotation. Words and language are not proper means for explaining and representing affective states as they are of different categories. This line of thought is further elaborated by Panksepp:

> "Affects fill the mind with a large variety of desirable and undesirable experienced states that are hard to define objectively or to talk about clearly. Partly this is because *raw affects* are pre-propositional forms of consciousness comprising brain and bodily processes of kaleidoscopic complexity." (Panksepp, 2008, p. 48)

So affective states are non-linguistic, pre-prepositional, which does not exclude them being intentional, and thus should be described in terms of neural systems involved.

Neurological accounts are similar to dimensional accounts in that they (i) recognise the immediate impact of affect on behaviour, preceding cognitive appraisal, and (ii) that they subscribe to the principal properties of affective states that can have different subjective feeling characterised by valence and arousal. However there are important differences between both accounts on these two issues.

First of all, (ad i) neurological accounts are more specific on the relationship between the stimulus and the response. Ledoux discovered for the fear system in rats that he studied in depth, later confirmed for humans, that there are two pathways that can mediate a conditioned fear stimuli: (i) the fast one going from sensory thalamus directly to amygdala that is responsible for eliciting emotional fear homoeostatic and behavioural responses and (ii) the slow one which includes sensory cortex on the way between sensory thalamus and amygdala, which allows for cognitive assessment of the stimuli and adequate adaptation of the behavioural response to the outcome of this assessment in the cortex (LeDoux, 2000). So how human brain responds to a conditioned fear stimulus is both the automatic fear response that is afterwards modulated by sensory cortex, so cognitively assessed. This proves the appraisal theories are wrong in which comes first emotion or cognition but are right about the general principle that cognitive appraisal of affect eliciting stimuli is carried out in the brain, and both normally takes place simultaneously yet the automatic hard-wired response is faster. Neuroscientists suspect that what has been empirically verified for fear system may likewise hold true for other emotional systems, which needs yet to be studied.

Secondly, (ad ii) cognitive neuroscience predominantly sees affective states, at least in some part, as natural kinds, i.e. discrete phenomena that correspond to excitation or inhibition of particular neuronal systems, and a single system corresponds to a given emotion kind. It is the following statement by Panksepp which Barrett chose as the starting point for his offensive on the believe that there are "kinds of emotion with boundaries that are carved in nature" (Barrett, 2006):

> "Until demonstrated otherwise, it is assumed that these systems constitute the core for the "natural kinds" of emotion." (Panksepp, 2000, p. 143)

So dimensionalists reject the view that specific emotional kinds could be separated from the plethora of affective states, meanwhile neuroscientists tend to agree that there are brain systems that correspond to particular affective or emotional kinds for which reason one can distinguish and objectively compare across individuals and even species different affective states.

Before we give examples of taxonomies of affective states under neuroscientific accounts of emotion let us come back for a while to the terminological discussion.

As cognitive neuroscience operates with more objective, neurological, categories when talking about emotions, it is easier for this field to come up with more sharp definitions of different affective phenomena. Let us rely yet again on Panksepp:

> "I use the term *emotion* as the "umbrella" concept that includes affective, cognitive, behavioral, expressive, and a host of physiological changes. *Affect* is the subjective experiential-feeling component that is very hard to describe verbally, but there are a variety of distinct affects, some linked more critically to bodily events (homeostatic drives like hunger and thirst), others to external stimuli (taste, touch, etc). *Emotional affects* are closely linked to internal brain action states, triggered typically by environmental events. All are complex intrinsic functions of the brain, which are triggered by perceptions and become experientially refined." (Panksepp, 2005, p. 3)

Such understanding of emotion and affect dubs our terminological conclusions in chapter 2, such that emotions are hard-wired programmes that involve a wide range of mind-bodily responses to an external stimulus. Emotional affects are similar here as emotional feelings or feelings of emotions that Damasio talks about which we referred to in chapter 2. Still there is a subtle yet important element in the above presented definition that needs further attention. It is that Panksepp makes a clear distinction between emotional affects from all other possible affective states that may arise as consequence of bodily sensations involving sensory pleasure or satisfaction of different intensity. However he underlines that all affects are brain functions, albeit not all are mapped already so he talks about those that he and other neuroscientists have studied. It may be however that the nature of emotional affects is that of natural kind meanwhile there are some other affects that nature is close to the dimensionalist account. Until these are studied in depth neurologically the dimensionalist accounts will remain theoretical speculations. We must emphasize that these speculations nonetheless provide useful generalisations and are indispensable to application of emotion theories in other fields like information science.

So cognitive neuroscientists propose different emotional affect taxonomies. The classical one was proposed by Ekmann including six basic innate emotions: *joy*, *distress*, *anger*, *fear*, *surprise*, and *disgust* (Ekman, Friesen, and Ellsworth, 1982). Panksepp's more contemporary account distinguishes a different set of *core emotional affects* that reflect transdiencephalic "energetic" action systems such as: *seeking*, *fear*, *rage*, *lust*, *care*, *panic*, and *play*, and defines each them in neural terms (Panksepp, 1998). It is important to highlight that brain scientists agree that there is a hierarchy of affective states such that some are *primary*, *core*, or *basic* emotions that are innate and which provide background for constructing higher level emotions that may be cognitively derived, learned and thus culturally determined. There is however a general agreement that as the basis for all affective states is provided by the core

affects every affective state must have an innate component, so that we should rather consider how innate a given emotion is rather than wondering if it is innate at all. For a historical overview of the topic please see Evans (Evans, 2001). This presents a potential direction for unification of dimensionalist and neurological accounts. It could be advanced that the dimensions are defined by core affects meanwhile the higher-level affective states derived from the core affects, cognitively and culturally acquired and strengthened. The variety of combinations of activation of different core affect systems could be conceptualized and explained as n-dimensional affect space where n is the number of identified core affects.

As we perceive experience as affectively coloured knowledge the natural direction for searching relevant formal systems apt for representation of experience is combining the approaches of knowledge engineering with those developed under affective computing. To our knowledge there are no systems that would undertake to represent emotions and knowledge on equal footing under a unified framework of experience representation. This is largely due to the domination of cognitivist standpoint which integrates emotions and feelings into the cognitive processes to which standard KR techniques are applied.

Importantly however, meanwhile the available frameworks can be applied to handling emotions, in so far as they correspond to the programmed behavioural responses, they ignore both immediate impact of affect on behaviour as well as emotional feelings, the by-product of emotions, that have a profound and long-term impact on cognition and behaviour. Although virtually all computational models of emotion include a component responsible for estimating the behavioural and cognitive consequences of affect, typically their role is limited to modulating consecutive behavioural response and reasoning, which usually takes for of *if... then...* type of rules that focus on short-term impact of affective states on behaviour. It is uncommon that such frameworks include a component responsible for long-term experience estimation necessary for emulating affect impact on behaviour that is deferred in time due to memory.

Consequently there appear to be two main directions for bridging the gap: (i) adapting the existing formal models of emotions and complementing them with experience estimation component, (ii) devising a new model, built from scratch relying on the revised theoretical foundations which incorporate recent insights into the nature of human emotions and feelings provided by brain research. Here we advance a new framework for emulating and representing experience in formal systems meanwhile we build on the improved existing computational models of emotion to represent the affective component of experience.

In conclusion, in this section we have reviewed the mainstream theoretical models of affect that provide the starting point for AI and information scientists for advancing emotion representation approaches and computational models of affect. The existing computational models of affective phenomena are unsatisfactory as they rely on

outdated appraisal accounts of human emotion. They evidently fail to embrace the complexity of affective dimension of experiencing by a natural agent, regardless human or animal. Evidently the difficulties in designing a sound and complete computational model of emotion and experience stem from the lack of complete and sound theory of emotions in first place. Most of the mainstream theories of emotion and experience that provide the foundations for the computational models has been refuted by the recent developments in the brain science, or have never been compatible with the earlier known facts which were simply ignored or skipped for the sake of provisional yet satisfactory conceptualisations. Still, the problem is not trivial. The difficulty lies in the fact that the computational implementations are driven by applications that require end-to-end solutions. No incomplete, half-frameworks are acceptable. The framework must be complete to serve its purposes and, unlike in brain science, it is hard to identify tiny sub-problems on which research effort could concentrate, gradually contributing small yet solid building blocks for the all-encompassing framework. In cases which these kind of tiny elements, treated as sub-problems, can be identified the progress is much more satisfactory. A good example are mappings of emotional states to facial expressions in the research on virtual characters or cyberspace avatars. Armed with sound theories worked out in course of decades-long research conducted by Ekman (Ekman, Friesen, and Ellsworth, 1982) and followers computer engineers could design robust and highly believable emotional virtual faces that aptly emulate human facial emotional expressions, which are vital for inter-human communication, shaping social relations and shaping culturally-driven emotions.

Importantly however, most of the models fail to propose algorithms that would link emotional responses to behaviour in a satisfactory way. There is currently no framework that could reliably predict human behaviour in affective situations. The problem still is in the lack of sound theories of human behaviour embracing both affect and cognition, and the problem of free will discussed in end of chapter 3 stands out still as an unsolved issue. Still the need for more accurate emotion representation frameworks is vivid, with opportunities in application to areas such as computer games, virtual and augmented reality, information and knowledge management systems, learning and computer-human interaction.

In the following section we will advance a general purpose framework for experience representation embracing cognitive (intentional) and affective dimensions of conscious human experience. In chapter 5 we will review selected computational models of emotion applied in the frameworks dealing with rational artificial agents. This area of research directly focuses on mapping affect onto rational behaviour for which reason these frameworks has been selected for further analysis.

4.3 Formal definition of experience

4.3.1 Experience modelling in the context of customer decisions

Our earlier work on experience representation and modelling in information systems appeared in the context of customer experience management. We tried to understand and quantify experience that results from the interaction with and consumption of goods, services or related entities such as symbolic goods: brands, and provide for a theoretical framework that would allow store information about affective aspects of customer experience. This earlier work will be briefed in this section.

In (Kaczmarek and Ryżko, 2009), considering experience in the context of customer behaviour, we have proposed that the experience be defined as remembered states of mind resulting from appreciation of stimulus events that determine generically any human behaviour. In the case of a consumer decision it was proposed that customer experience could be modelled as a set of learned concepts about an object of consumption (brand, product, service, provider, etc.) internalised by a given customer. It has been proposed to look at experience gaining as a learning process, and treat transactions and consumption related events as training examples. Two years later in (Ryżko and Kaczmarek, 2011) we proposed how to further develop the above idea and translate it into a more formal model. We first narrowed down the definition of customer experience slightly by exchanging the remembered *states of mind* with remembered *intentional states* resulting from appreciation of stimulus events related to a consumption object, which allowed us to represent these states with predicates, and apply predicate calculus to experience processing. Furthermore we attributed experiential contents (the predicates) with emotional valence and intensity, noting that the predicates representing the learned concepts about an object of consumption are of "belief" type and have temporal dimension.

As far as the customer experience was concerned the psychological mode of the intentional state was a *belief* and propositional content was a predicate of argument x, where x was a consumption entity. On top of that we modelled the emotional value of an experiential intentional state as a variable that could be a number or a logical value represented as an attribute of the given predicate. Consequently we could formally define customer experience as follows:

$$Exp(t) = \{< p, v >: p \text{ is a predicate believed to be true at time } t, \text{ and } v \text{ is its emotional value}\}.$$

The critical point in our approach was that the experience gaining was modelled as learning process so that theory of machine learning could be applied to modelling accumulated experience change in time. As any learning process involves training data, which goes through the learning algorithm and results in a set of learned

concepts (intentional states), in the context of customer experience we considered a set of events involving a particular consumption entity and the particular customer as training data, duly ordered in time. Consequently events were represented as a tuple of the following format:

$e =< d, t, c, v >$, where d is event description, t - time, c - class (e.g. offer, advert, sales etc.), v - vividness

The actualization of accumulated experience in turn was proposed to be governed by a learning function that processes the above defined training data (events) into an updated state of total experience (a complete set of all predicates) by generating new experiential intentional states as outputs and/or altering the experiential intentional states generated in the past. Noteworthy, the proposed architecture was generic and agnostic in terms of knowledge representation formalisms and learning algorithms. The universality of the approach was demonstrated by applying a selected defensible reasoning formalism, to be more precise default logic defined by Reiter (Reiter, 1980). Default logic allowed us to model the customer experience actualisation over time and customer deliberation in customer choice. In the former case the model benefited from the non-monotonicity of the chosen system, in the latter we took advantage of the feature of default reasoning that allows the reasoning process to end up in multiple possible worlds as to represent the psycho-cognitive fact that people confronted with choice tend to construct different concurrent alternatives and weight arguments for each of them, before finally committing to the chosen one (Salzman and Newsome, 1994). The argumentation phase of the deliberation process was presented as a moment where emotional valence of experiential instances could step in and impact the final choice by influencing rule priorities in the model. Although the benefits of adopting default reasoning were clear other formalism were claimed to be applicable as well, which however remained unexplored.

In conclusion let us note the most important elements of the so far discussed formal model of experience: (i) it was developed with the purpose of being applied to representing and modelling experience in the context of consumer decisions which narrowed the model scope; (ii) despite introducing the concept of emotional valence of instances of experience its application to formal reasoning has not been fully exploited; (iii) although it has been demonstrated that default reasoning could be an appropriate formalism to be applied to practical implementation of the model other alternatives are still unexplored as well as it remains open whether the default reasoning is the ultimate choice, in the sense that it can meet all the requirements imposed by the model; (iv) relation of experience to time has been addressed however implementation details remain widely open, both in terms of the general forgetting algorithms as well as the influence of emotions on experience consolidation; (v) finally and most importantly in its foundations the model entirely depended on the mainstream knowledge representation formalism the predicate logic and only

superficially takes account of the affect, which is as we tried to demonstrate earlier a component of experience that is an integral and indispensable complement to knowledge.

4.3.2 Towards a general purpose definition of experience

In order to be able to advance the above presented approach to representing customer experience towards a general purpose framework for human experience representation in information systems certain generalisations and modifications must be introduced.

We start with a broad understanding of experience as remembered intentional states of mind. Formally, experience can be defined as a pair of sets K and A, where K represents knowledge, that is contents of remembered intentional states of mind, or intentional contents of experience, whereas A represents affect, i.e. the subjective qualitative component of experience, therefore:

$$E_J =< K, A >, \text{ where, } E \text{ - experience of agent J.}$$

Further we define a funcion mapping intentional content into affective state

$$f : K \rightarrow A$$

It is important to highlight the following on the elements of the above proposed definition. Experience is intrinsically a subjective phenomenon, so it only makes sense to talk about subjective experience, therefore an experience of a given agent. Experience exists in an individual mind and for this reason it is impossible to define experience in relation to *episteme*. Talking about knowledge as a component of experience we will always talk about *doxa*. Episteme is an abstract construct and as such cannot be described in first personal narrative, Tarski's concept of truth supports this as requires meta-level language for being able to define truth in first place (Tarski, 1933). Episteme can be experienced, but then becomes *doxa*. As knowledge matters only when it appears in a conscious mind, ultimately it always takes a form of *doxa* and it is never false as such, merely is composed of propositions that are either true or false, because anything to be knowledge would require to be in a mind and by mere fact of appearing in a mind it becomes knowledge even if is in objective terms an illusion.

K – knowledge, is defined here intuitively as *contents of intentional states*. K in principle can be represented with any known method of knowledge representation as long as it complies with the above definition.

A - affect, represents the subjective feeling accompanying the occurrence of the intentional content in the consciousness. As discussed in chapter 3 the indispensable element of consciousness are the primordial feelings, or feelings in general that are coming from the entire body. These are the Damasio's "feelings of what happens" which provide the background scene for intentional contents appearing in the mind.

These feeling states constitute A. Anatomically these feelings may be a compound of various neuronal signalling coming to the brain but they become unified in the field of consciousness so we can treat A as a single entity. However for the purpose of description and representation of A we may distinguish some separate *properties* of A which is why it is tempting to to represent A as a vector space which dimensions are defined by these properties, which we will elaborate further on. An element of A that affects the intentional state is an integral part of experience instance, as there can be no intentional state without affective dimension [1].

Consequently we are using the word *affect* in the same sense as Panksepp used it to coin the term *affective neuroscience* to mean the nature of our positive and negative experiences, i.e. the many ways we can feel about things instinctively. Affect will embrace under the proposed model all kinds of affective states: core affects, feelings of emotion (emotional affects), bodily or sensory feelings. We assume, following neuroscientists, that each affective state has its neuronal correlate, either in form of a single system, as in the case of core affects, or in form of a combination of different level of activation of different system.

We further accept that each affective state can be at least characterised by three properties: V - valence, i.e. either a point on a continuum between eternal pleasure and eternal displeasure, or a given value from a discrete set of values, as we know from both human and animal research that natural agents can always unambiguously discriminate between affective states they like or dislike (Panksepp, 1998), I - intensity as affects of the same valence can be stronger or weaker and natural agents show preference between two states of the same valence, M - mode, which correspond to a given affective kind, as we adopt the neuroscientific account that core affect are "natural kinds", consequently the set of values of M is composed of labels that correspond to either a given core affect or a derivative affect being a combination of core affects. So the element representing affective component of experience – A can be defined as a set of triples:

$A = \{< v, i, m >: v \in V, i \in I, m \in M\}$, where V – valence, I – intensity, M – mode, and M is a k-combination of C where C – n-element set of core affects and $k < n$ – the number of core affects involved in a compound affect.

To sum up, the affective component of an experiential state is characterized by *valence*, as a mindful organism can always discriminate between wanted, unwanted or neutral subjective states, *intensity* as there can be degrees to which these states are wanted or not and finally they are characterised by a *mode* as there are neurologically recognised emotions each corresponding to the activation of a particular neural circuit

[1] It remains an open question if there can be feelings without knowing. Very likely that some animals (Panksepp, 1998) and people with lack or serious lesions of cortical parts of the brain footnote may have such experiences. The case of a child Nicholas Coke who was born only with a brain stem fully developed, or other children suffering anencephally or hydroencephally discussed in Damasio (Damasio, 2010) may serve as clinical cases for further studying of this kind of experiences

in the brain, or a few circuits at a time. In line with the contemporary neurocognitive theories of emotion and affect (Panksepp, 2005) we distinguish between low-level, primordial affective states, i.e. *core affects*, like fear, lust, etc, and compound or high-level affects that can involve a combination of core affects, which is why we have defined M - mode, as k-combination of the set C, where k is any integer such that $k \in < 1; n >$. Noteworthy, we do not take into account to what degree each particular system corresponding to an element of C is activated. We judge valence and intensity for the elements of M only, which is in line with the account of the affective component of experience as unified and subjective, and subjectively such nuances cannot be consciously appreciated. We consider it a fair and accurate approximation.

The reason why we insist on including M as important parameter is because we assume that different affective kinds can be mapped onto behavioural tendencies that these affective states imply. If one would want to infer on the likely behavioural response by an agent to a given stimuli not only intentional contents captured by the mind during the event in question, its affective value in terms of valence and intensity are needed but also the kind of affective state that has been provoked by the event, as to equal situations in terms of K, V and I may result in different behavioural response under different M. Evidence to this claim can be recalled from experiments in neuroeconomics. Knutson and Winkielman for instance have shown that in the same consumer decision situations prospects are less likely to buy an offered product if the insula, part of the brain that mediates sensory affect of disgust is activated, similarly prospects are more likely to choose a risky option after being exposed to a positive (happy) facial expression than after experiencing a negative facial expressions like anger, disgust or fear (Knutson et al., 2007; Winkielman et al., 2007).

The intentional states of mind, which build up E, carry contents (K) that can be represented with prepositions of natural language which have a logical value: true or false. The propositions themselves only carry intentional contents but do not carry affective contents, still they are affectively marked. Each intentional state of mind, as it exists in consciousness is always accompanied by a subjective, affective feeling of a living body that constructs this consciousness. The intentional contents do not carry the affective contents rather they are marked, in the sense that they have capability of invoking affects of a particular type. The affective component of intentional stance is not a stable entity but a highly epiphenomenal entity. In consequence of the memory processes in the brainmind, the affective component of intentional state is always reconstructed whenever the given state appears in the mind and the end result of the reconstruction process depends on past experiences, i.e. the remembered affective marking and the current experience that has a given feeling to it. The reconstructed affective state mingled with the intentional state in the unified field of human consciousness becomes thus the updated affective marking of the intentional state and rerecorded in memory for future reconstructions. This process

can be looked at as modification re-composition of patterns which is influenced by new knowledge and in parallel new affective states.

Furthermore experience exists only in the present, albeit it depends on past experiences which together with emotional reactions that are biologically determined provide the building blocks for the reconstruction and composition of the present experience. Therefore each instance of experience is reconstructed from the remembered and current experiences. Consequently the experience has a clear temporal dimension but the character of this dimension is not that of a linear, time-ordered sequence of events but rather is defined by vividness of remembered past experiences. Due to the unified nature of experience the past remembered experience is always updated whenever recalled and reconstructed from memory. So the recall of a remembered experiential instance is reshaped with the current one and rerecorded in the updated form. Furthermore past experiences decay with time if not recalled and reconsolidated on a regular basis. Here algorithms emulating the process of forgetting such as proposed by Woźniak (Wozniak and Gorzelanczyk, 1994; Wozniak, 1995) could be applied.

In the next section we will advance a fully-fledged experience representation framework by Kaczmarek and Ryżko, based on the above formulated account of experience.

4.4 Kaczmarek-Ryżko framework for experience representation in information systems

The Kaczmarek-Ryżko Experience Representation Framework (KRERF)[2] starts with the general purpose definition of experience as proposed in previous section:

$$E_J =< K, A >, \text{ where, } E \text{ - experience of agent J.},$$

and a function f mapping intentional content into affective state:

$$f : K \to A,$$

where the element representing affective component of experience – A is defined as a set of triples:

$A = \{< v, i, m >: v \in V, i \in I, m \in M\}$, where V – valence, I – intensity, M – mode, and M is a k-combination of C where C – n-element set of core affects and $k < n$ – the number of core affects involved in a compound affect.

[2]A paper presenting the latest version of the KRERF by Kaczmarek and Ryżko has been accepted to the 2012's International Conference on Brain Inspired Cognitive Systems (BICS 2012) and will be published in Springer's Lecture Notes in Artificial Intelligence.

Furthermore KRERF emphasises that the experience is gained over time in course of agent's interactions with the environment. Each such an interaction for us is an *event*. For this reason, as earlier, we take experience gaining as a learning process to which machine learning approach can be applied. Any learning process involves training data, which goes through the learning algorithm and results in a set of outputs that are learned concepts, in our case these concepts are experiential intentional states, elements of E_J. Consequently, in experience gaining process we will consider *events* as training data. For the purpose of memory modelling, including memory decay and consolidation, we will need to control the time line of events for which reason more formally define it in the following way

$$T(t) = \{(e_1..e_n) : \forall k \in (1,n)\ time(e_k) > time(e_{k-1})\ and\ time(n) <= t\}, \text{ where}$$
$$time(e) \text{ is the time of occurrence of } e$$

Each event is a tupple:

$$e =< B_e, t, a >, \text{ where } B_e \text{ - believes resulting from the event } e \text{ (a set of intentional}$$
$$\text{states to be remembered), } t \text{ - time, } a \in A \text{ - affective value of the event as defined}$$
$$\text{earlier.}$$

It is important to note that we assume that affective value is assigned to an event and inherited by B_e, however this value may result both from the affective response to external stimuli prompting the event, as well as mind processes caused by the stimuli, i.e. reconstruction of affected past experiences. This is to embrace a situation when an agent involved in cognitive response to a stimulus recalls strongly affected facts and this sets the affective value for the event and consequently the new output believes resulting from the event. It is important because so derived affective value of an event a can reach a certain threshold level such that it may trigger an autonomous behavioural response instead of a deliberative response, i.e. a fully fledged emotional response, which intervenes a regular deliberation process and may result in behaviours that are irrespective of cognitive appraisal of the event. This calls for defining a mechanism for updating a both at the level of event as well at the level of believes B.

So far we have focused on E and A, let us consider K briefly now. As we suggested earlier, principally K could be represented with any KR approach. As K is composed of intentional states of propositional nature the elements of K can be represented with simple logical propositions, or predicates, so as truth/false-valued formulas of a formal language, with or without variables, where logical value is determined by mind-to-world relationship. In KRERF we favour predicate calculus for representing agent knowledge and reasoning. However as we argued before K is not enough to represent comprehensively the state of mind of agents, as it lacks the subjective, affective component of experience. Consequently, while considering intelligent agents,

BDI agents in particular, their knowledge base, B – believes set in case of BDI, should not be limited to K but should be assumed E.

Consequently, under KRERF using predicate calculus we can represent experiential intentional states, remembered by an agent from the event, as a set

$E = \{p_1(a_{11}, x_{12}, ..., x_{1m}, c_1, a_1), ..., p_k(x_{k1}, x_{k2}, ..., x_{kl}, c_k, a_k)\}$, where $p_1 ... p_k$ are predicate symbols and $x_{11} ... x_{kl}$ its attributes, while c_k and a_k are *consolidation coefficient* and *affective value* of a predicate p_k.

Variable a_k is derived and updated dynamically during each event from the affective value of the current event e as well as affective value of remembered past experiential states in which $p_k(x_{kl})$ appears. Moreover, the variable a_k should be also dependant on affective value of *related* affected predicates, which relation could be determined by co-occurrence of certain events or based on associations between contents of the predicates. For the purpose of a_k estimation an *affect update function* should be defined. As there are only general premises on how such a function should be build, we define it on purpose in a general way so that a more specific implementation can be used for a particular application.

Variable c_k is derived taken into account the time of the current event and the time of the past events in which the same predicate was remembered. For the purpose of c_k estimation a *decay function* must be defined. This function should reflect the fact that experience consolidation depends both on time, repetitive recall of the given predicate from the memory, as well as the role of affect in the memory processes (Kandel, 2006). For defining the decay function one could build on available memory models and algorithms such as Woźniak's algorithms based on spaced repetition (Wozniak and Gorzelanczyk, 1994). We insist that the temporal dimension is one of the most important aspects of the experience. As an agent is confronted with new events, the experience gained from the old ones will be steadily forgotten. As forgetting is intrinsic to learning the model of experience must take this into account. So the decay function governing the vividness of past experiences, represented by consolidation coefficient (c_k) must be provided for to enable modelling of the memory volatility linked to experiential outputs of events.

4.5 Conclusion

The evidence from both contemporary philosophy of mind and neuroscience discussed in previous chapters seems sufficient to claim that affect play a central role in human and animal decision making (Damasio, 1994; Panksepp, 1998). Based on this evidence let us postulate that any intentional state, which is as explained earlier representation of external world in human mind, has an affective value, which is characterised by valence (positive or negative), intensity (arousal level) and mode (affective state

kind), which has implications on agent's behavioural response and is an integrated component of agent's rationality.

The model for experience representation we propose challenges the mainstream affective computing accounts of emotion representation is so far as intelligent, rational artificial agents are concerned. As we speculate that our take on representing affect and affective experience in artificial agents, for being more in line with contemporary account of consciousness, affective neuroscience and rational agency, is likely to outperform currently available approaches in emulating natural agent behaviour in information systems,

Next chapter will consider experience in the context of agent's rational choice. Our ultimate goal is to explain how our account of experience and the proposed experience representation framework could enhance the mainstream approach to modelling purposive behaviour, thus deliberative decision making by natural agents. We will also undertake to explain why we believe our approach to experience modelling is superior, especially in the context of rational, deliberative agency.

Chapter 5

Application of the framework to modelling rationality of an experiencing agent

The purpose of this chapter is to theoretically verify the applicability of experience representation model proposed in previous chapter to emulation of rationality in information systems. This chapter will look into how the proposed model could enrich mainstream agent rationality frameworks by embracing agent's experience as decision variable.

The fundamental question that must be answered here is why such a model is needed in first place. The natural hypothesis is that the existing models are unsatisfactory, which must be verified. The next important question that follows is how adding the emotional dimension to agent's rationality model can improve models efficiency. Finally a practical way of including this dimension in the models currently used by information systems designers should be devised.

To accomplish the above this chapter will first present the classical rational decision maker model, commonly referred to in classical economics literature as *homo oeconomicus* or *economic man*. Later a body of research delivered over last 50 years within many disciplines that undermines the foundations of classical model will be presented. This will embrace the early criticism on the philosophical and ethical grounds, the main wave of criticism from the behavioural psychology reported throughout the second half of XX century, as well recent insights from cognitive sciences, including brain sciences and philosophy of mind. However this review has not been structured with the purpose to document the line of attack on the classical model, rather to provide an overview of important findings of various disciplines that has influenced importantly the evolving model of rational agent and that are still impacting it and stirring discussions on its relevance. Primarily however this will be done as to identify major weaknesses of the classical model that could be addressed by the experience representation framework we propose.

Next, I will present an overview of what are the contemporary mainstream approaches to rational agency, that are to certain extent the outcome of the questioning of the classical model, and what are the strategies for unification of these theories across many disciplines. This will allow us to conclude that in information science and IA the mainstream approach to modelling rational behaviour is that proposed by Bratman (Bratman, 1987; Bratman, 1999), involving agents that organize their behaviour with *believes*, *desires*, and *intentions*, the so called BDI model.

Then we will look at the representation methods of rational agency in information systems, with emphasis on the mainstream approach to modelling artificial rational agents, including BDI agents that are engineered with different formal BDI frameworks, as to provide a brief review of the ways the decision making can be represented in information systems.

Next, we will consider how the experience representation framework proposed in previous chapter could be utilised for enhancing the current methods of agency emulation in information systems. To this end we will provide an overview of the state-of-the-art in emotional agents formalizations and finally suggest a way for integration of our experience representation framework with BDI approach to modelling agent rationality. Consequently this chapter will set the foundations for emulating behaviour of experiencing agents

5.1 The Classical Model of Rationality - TCMR

5.1.1 Towards TCMR account of rationality

What is rationality, or what being rational means is not always obvious. An important deal of misunderstandings and scientific disputes on the topic of rationality stems form imprecise definition of rationality. In this chapter I will primarily focus on rationality as understood by economists and behavioural scientists, as more philosophical and ethical perspective on rationality has been provided in chapter 3. As for this thesis rational decision making, hence the rational behaviour of a human being is pivotal, let me adopt a general yet inclusive and useful definition of rationality as acting consistently according to a certain set of rules. It is important to recall however that rational behaviour is sometimes defined more narrowly, by attaching a certain qualitative value to rationality, in other words acting rationally means acting in some qualitatively specific, acceptable way. This is achieved by limiting the domain for the set of acceptable rules, by proposing that these rules of conduct are of some particular character. For instance that these are rules dictated by 'reason', 'morality' or that these rules are shared by a group of individuals, by which fact a rule constitutes a social norm. Such an approach to rationality is strongly driven by historical disputes over the term, which is central to many, often distant disciplines.

The understanding of rationality as 'consistency' in behaviour is not new whatsoever in economics. Lionel Robbins discusses this in his essays published in 1932 while commenting on von Mises's theory of human action, noting that choices made by people should not be divided into rational and irrational using these terms with a normative significance. Instead a conception of rationality as equivalent to consistency should be used, understood in a wider sense as figuring in discussions of the conditions of equilibrium (Robbins, 1932, p. 92).

Importantly, to speak about rational behaviour we must define the acting subject, which we have earlier agreed to call an agent. An agent that acts consistently according to a given set of rules is therefore a rational agent. Given such a definition of rationality, a model of rational agent would have to specify in first place the set of rules, as well as the set of meta rules, i.e. what primordial processes determine the set or rules, and/or a set of axioms, i.e. the rules that are taken for granted without justification. Such a formulation of the problem naturally suggests that a representation of rationality in artificial systems could be possible on the grounds of formal logic. I will come back to this point in the consequent sections.

As already mentioned in the introductory chapter human behaviour historically has been scientifically dealt with by economics, which is why the most prominent and famous model of rationality comes form this discipline. This is the model of *homo oeconomicus*, or the *economic man*. Before I present and discuss this model let me reflect first on its relevance to the economics and behavioural science at large.

A human being in action can be seen from two slightly different angles: (1) as an agent that is constantly being confronted with choices which one is predestined to make, and one cannot avoid as inaction is one of the available choices; (2) an agent that conducts a purposeful actions that is aiming at ends and goals as expressions of its will. The difference is subtle but I claim this difference was pivotal for shaping differences in scientific perspectives on rational human action. Gintis somewhat confirms that when he suggests to distinguish deliberative decisions from routine decision making as way of approaching the unification of decision theory for behavioural sciences that he postulates (Gintis, 2009). Noteworthy, these two approaches do not have to contradict at all, on the contrary can complement each other, still the divergence of these two approaches has led to quite opposing and competing models of rational agency. The former approach is closer to neoclassical economists, game theorists, subjective expected utility theorists and behavioural economists such as Edwards, Savage and Gintis, for who the 'consistency' aspect of rationality is pivotal, meanwhile the latter to behavioural psychologists, cognitivists and philosophers such as von Mises, Simon or Searle, who put more emphasis on purposive character of rational action. Noteworthy purposiveness, in other words goal or ends orientation is considered by the latter group as constitutive element of rationality.

The meeting point for both approaches is the choice itself at which an agent arrives in one way or another. The reason why rationality has become so important to economists is that choice is tightly bound with scarcity which is a central problem for economics (for details see Chapter 2). Choice is the consequence of scarcity, or as Barron and Lynch put this the problem of choice arises only if there is scarcity. As Hall and Lieberman rightly observed the scarcity of resources and the choices that it forces people to make is the source of all problems studied in economics (Hall and Lieberman, 2001) but also in other behavioural sciences. As discussed in chapter 2 choice can be defined broadly or narrowly. In the narrow meaning of the term choice is an indication of one of the options. In a broader sense, much more common in economics and psychology, the choice is a conscious mental process that results in the selection of one of the available options by comparing their utility to the decision-maker. The choice in a wider sense is therefore a complex process involving: identification of available options, their assessment, comparison and finally an indication of one of them. If an agents acts upon the chosen option we can talk about a decision in a proper meaning of the term.

Therefore a body of economic theory concentrates on choice optimization, in particular as far as choices made by consumers and producers on the market are concerned. The part of economics dealing with choice is simply referred to as consumer choice theory (theory of consumer's choice), also called the theory of rational choice (rational choice theory) (Edwards, 1954), which relates to consumer behaviour in the market and their decisions regarding consumption of specific goods and services.

This is where another aspect of classical rationality comes in: *optimisation principle*. A rational agent is supposed to be consistent with applying rules, but according to the classical model of rationality the most fundamental rule is that an agent prefers more than less, so that his rationality dictates what and how should be optimised, or using the concepts introduced so far, what a rational agent should be consistent about. The issue of choice optimisation has two facets: (1) how choice should be made, i.e. how to make choices efficiently; (2) how people do make choices and how this impacts the micro and macro dynamics of the economy. The former approach is typical for normative or prescriptive economics, the latter for positive or descriptive economics. It is important to distinguish between these two approaches as they imply distinct objectives and provide different kinds of theories and insights although in principle they operate on the same theoretical basis, i.e. the notion of rationality, choice, action, means and ends, agency, etc.. This thesis in particular focuses on the positive or descriptive choice theory, therefore how people do behave, not ought to behave, when confronted with choice. This is because the primary objective of this thesis is to contribute to the improvement of the models that simulate and explain consumer choice observable on the market.

5.1.2 Homo oeconomicus model

The understanding of rationality has been highly influenced by the classical economic model of rational decision maker: *homo oeconomicus* or *economic man*. For nearly 200 years, since the times of John Steward Mill (1806-1873), the economic man model has been dictating what rational has meant, not only in science but also in everyday speech.

At this point let the basic assumptions of the model be introduced. There are three main suppositions that define the economic man: (1) he is completely informed; (2) He is infinitely sensitive; (3) He is rational, in the sense that (3a) he can weakly order all states of universe that can be available to him by which he defines his preferences, (3b) always chooses the best of them as he always maximises the value that he can expect from the choices available to him. For completeness it must be added that the preferences of economic man must be transitive, which means that given available options A, B and C, once he prefers A to B and B to C he is expected to prefer A to C (Edwards, 1954). Unsurprisingly, the latter requirement regarding economic man's preferences is one of the main targets of attacks by the opponents of the model along with the second controversial axiom that says that rational behaviour must imply maximizing something. The arguments against plausibility of the first axiom will be discussed in detail in the following section that will cover an overview of empirical and theoretical research results in this respect.

As for the maximization problem the economists had to answer a non-trivial question: what is to be maximised. For this purpose they introduced a highly useful yet quite vague and, with years, increasingly abstract concept of utility. Very likely the birth of the abstract concept of utility was dictated by trying to avoid entering into moral discussions on the model. Economists often underline that their methods are morally agnostic, following the postulate by Max Weber that social sciences, which even though may require explanation of conduct that involves the consideration of certain data that are not purely of physical character should be freed from subjective evaluations (*wertfrei*). Noteworthy, the initial criticism of the classical model of economic man's rationality spearheaded by John Ingram (Ingram, 1888) was held on the grounds of ethics. The *homo oeconomicus*, which term Ingram coined in his classic *History of Political Economy* was attacked as a false archetype of a human being that implies people are extremely selfish and rationally egoistic, i.e. acting primarily in their self-interest.

Indeed the concept of utility as well as its conventional unit: *utile* comes from utilitarians, Jeremy Bentham (1748-1832) and John Stuart Mill (1806-1876) who have seen the greatest moral value in the behaviour that leads to the greatest happiness for the greatest number of people. Utility was in the beginning associated with sentiencial feelings or happiness or pleasure, and disutility or negative utility with pain or suffering. However as suggested earlier the concept of utility has evolved towards a

more abstract idea of preference utility which means satisfaction of preferences. This evolution took a few steps and was started by Edgeworth (Edgeworth, 1881) who assumed, unlike early utilitarian theorists, that utility was not an additive function of utilities of separate independent goods, as different commodities can not be combined into a total utility by simple addition. As a consequence Edgeworth came up with the notion of indifference curves representing compounds of goods that are of the same value to a consumer switching between which resulted in no change in consumer's utility. This way the utility concept that was associated with some absolute value measured with the amount of pleasure or pain was replaced by a more abstrct concept of utility that represents relative value of different bundles of goods. Further, in course of work of Pareto (Pareto, 1906), (Pareto, 1971), Johnson (Johnson, 1913), Slutsky (Slutsky, 1915), and finally Allen and Hicks (Allen and Hicks, 1934) and Wold (Wold, Shackle, and Savage, 1952) the modern classical theory of riskless choice has been shaped where the n-dimensional commodity space is considered, with n-1 dimensional indifference hyperplanes in that space. This theory was equipped with an elegant mathematic structure based on linear algebra and calculus, with allegedly a minor flaw: a formal constraint that consumers shall always have a complete weak ordering for all commodity bundles, or points in commodity space (Edwards, 1954). To complete the above picture it must be added that Samuelson observed that indifference curves, hyperplanes and entire indifference maps can be derived from empirical observations of consumers' choices (Samuelson, 1947). The process of making the concept of utility abstract has undoubtedly alienated it from how normal internal deliberation process of a real decision maker looks like. Undoubtedly most of non-specialist making choices do not think about indifference hyperplanes, nor even may know what utility means. This trivial observation has fuelled a massive criticism of the model mainly by psychologists, to which the most common response by decision theorist was that the model well describes the outcomes of peoples decisions, not the internal psychological processes of deliberation, which are irrelevant to economics. This argument is being used until today and will be taken up again in the following section.

To complete the presentation of the classical decision maker model it is necessary to add risky choices to the scope. The classical model has been extended to embrace risky choices principally by von Neuman and Morgenstern (Von Neumann and Morgenstern, 1944) in their classic publication *Theory of games and economic behaviour*. The main modification they introduced to the model was that economic man can also completely order probability combinations of states. This allowed for deriving cardinal utility as indifference between possible probability combinations of states, using the equation for expected utility: $EU = p_1U(s_1)+p_2U(s_2)+...+p_nU(s_n)$, where $p_1+p_2+...+p_n = 1$. The consequence of this approach, among others where that: (1) risky propositions can be ordered in desirability, just as riskless ones can, which allow to lesser the requirement of complete information; (2) the concept of expected

utility became behaviourally meaninglful; (3) choices among risky alternatives are made in such a way that expected utility is maximized (Edwards, 1954).

From the outline provided above a model of static expected utility maximising agent emerges. Importantly this model involves evaluation of subjective value or utility of possible outcomes and deriving their objective probability. In 1954 Savage introduced an extension to the model proposed by von Neuman and Morgenstern, which became a classical theory of choice under uncertainty called the *SEU* model (*Subjective Expected Utility*) (Savage, 1954). Mathematically the model has not changed at all compared to expected utility model, however the interpretation of the utility function was modified importantly. Savage assumed that expected utility does not depend on objective probability, indeed often the decision makers are ignorant about probability theory and ways how objective probabilities are calculated, instead their conduct is governed by their own subjective assessment of how likely a given event is. As a result the concept of subjective probability was born (though Savage called it originally personal probability). As a consequence the measurement of the utility depended under the SEU model on both subjective value and subjective probability. To keep the formal soundness of the model Savage had to introduce an assumption that subjective probabilities equally to objective ones are additive, so that probability of two mutually exclusive events is equal to the sum of the two probabilities. Early psychological experiments, which tried to directly estimate subjective probabilities with psychological methods under laboratory conditions, seemed to prove that this assumption is plausible, as they showed individuals' subjective probabilities are linearly related to observed proportion (Edwards, 1961). Another important principle introduced by Savage was the *sure-thing principle*, which says that if a course of action A is at least as good as course of action B in all future states of the universe, still A is better in one or more states, than B should never be preferred to A. In other words that the preference between two gambles a and b which have the same set of events and which have identical consequences in one of the events (but not necessarily in others), should not depend on what that identical consequence is. The sure thing principle allows to assume that probabilities of events are independent from the utilities of the outcomes of these events, which opens the ways for better exploitation of the model with mathematical apparatus and theorems of probability theory.

The sure-thing principle, additivity of probability and the transitivity of preferences constitute the three main pillars of the SEU model, which is used in economics still today. The widely used argument for the validity of this model is the *Dutch book* or *Duch lock* argument, which states that a decision maker constantly violating the transitivity or additivity principles would be eliminated from the market as it is possible to construct gambles (or games, trades) that are strictly worse to him which in longer run would lead the decision maker to sure loss (Freedman and Purves, 1969).

To conclude this introduction, according to the classical notion of rationality in economics, to which I will refer as The Classical Model of Rationality (TCMR), (1) the decision maker while confronted with choice always maximizes his utility (prefers more than less); (2) towards this end he can weakly order all the courses of action available to him in which (3) he is consistently logical, in the classical meaning of the term, which results in, among others, completeness and transitivity of preferences (mathematically represented as binary relations); (4) in case of uncertainty, economic man applies probability theory to evaluate the expected utility of available courses of action in which again he is logically and mathematically consistent so that the sum of estimated probabilities of the available options equals to 1 and the preference transitivity principle holds; for this to be feasible (4) he must be completely informed which implies he can identify all the courses of action available to him, as well as knows the corresponding outcomes, or at least can estimate the probability of their occurrence, so that expected utility of these outcomes could be derived.

The next section will present the selected critical arguments on TCMR mainly provided by empirical studies on human behaviour. The objective of section two as a whole is to provide insights from different disciplines that contributed, and are still contributing to the evolution of the rational agent model, and in some contexts its abolishment.

5.2 Limitations of TCRM - interdisciplinary perspective

The model of rational agent is one of the most important one, for economics but also increasingly for behavioral sciences, including psychology, economics, cognitive science, evolutionary biology and anthropology. Many proceeding theories and models, including a few that brought a Nobel Prize to their inventors in economics, rely on it. It is fundamental and axiomatic paradigm for economics. This solely explains why the model both raises controversies as well as attracts many supporters and is still widely applied, however currently the major flaws, or better limitations, of its classical form are widely recognised which is why it has evolved, and still is evolving towards either more subtle (less constrained) or complex formulations in modern applications. The continuous effort by economists aiming at systematizing decision making process often aroused opposition by psychologists, as Hogarth and Reder provocatively capture as the continuous struggle of economic theory against falsifying evidence provided by psychology (Hogarth and Reder, 1987). The main axis of disagreement goes along the axiomatic suppositions of the classical rationality model.

As mentioned earlier, the initial criticism of the classical model came from the part of philosophers of ethics. Ingram depicted the Economic Man as immoral creature,

that acts selfishly in achieving self-oriented goals (Ingram, 1888). Much later, in the mid XX century, the ground-breaking work by Herbert Simon, Daniel Kahnemann and Amos Tversky and other contemporary choice theorists spearheaded what may be referred to as the behavioural revolution in classical theory of choice. This work prompted a widespread recognition that the psychological predisposition of a human being are not well taken into account by the model of empheconomic man, as humans simply do not decide the way the classical model dictates.

The criticism of TCRM by the behavioural psychologists was particularly interesting compared as it confronted the model with empirical studies aiming at its falsification or corraboration, depending on initial hypothesis of a study. This empirical research followed the methodological line aptly expressed by Newell and Simon in their pioneering paper introducing GPS (General Problem Solver):

> "If we succeed in devising a program that simulates the subject's behavior rather closely over a significant range of problem-solving situations, then we can regard the program as a theory of the behaviour. How highly we will prize the theory depends, as with all theories, on its generality and its parsimony - on how wide a range of phenomena it explains and on how economical of expression it is."

In other words psychologists treated the classical model of rationality as a theory of human behaviour, which should be considered a good theory only if empirical tests involving real decision makers and real decision situations prove the predictions made by the model significantly accurate.

This section will recall the research results reported by critics of TCMR, in particular the ideas put forward by Simon's and Newell's as well as Kahnmann's and Tversky's and their colleagues as well as more recent evidence from philosophy of mind, cognitive neuroscience and neuroeconomics. This summary will take a form of a catalogue of what in literature is commonly referred to as *paradoxes*, or *biases* and *heuristics* in rational decision making, to which we will apply a general term *behavioural effects*, or simply *effects* for shortness. This catalogue of effects will continue throughout the chapter, therefore for the sake of ease of reference each one will be given a number.

As we present the behavioural effects, which basically pose a challenge to behaviour emulation in IS, we will already try to propose how our experiential approach to modelling rationality, in particular how the availability of information about subjective dimension of experience, could help facing this challenge. Importantly however, not all effects can be immediately eliminated by applying experiential approach, at least not in the way the effects are framed. For the purpose of completeness and consistency with literature we have decided to include an inclusive list of effects and in the way phrased and discussed in the literature of the subject.

5.2.1 Classic behavioural effects

Effect 1 *The St. Petersburg paradox*

The first behavioural critic of the classical stance that rational mean maximising expected value (currently utility) can be dated back to the famous paper by Daniel Bernoulli, published in 1738, in which he presented and explained the St. Petersburg paradox (Bernoulli, 1954). In brief, this essay tried to explain why people tend to avoid risk and why risk aversion decreases with wealth. Bernoulli constructed a gamble involving choice between a lottery and sure gain and confirmed empirically that people in general would prefer sure gain even if mathematically calculated expected value of the lottery was higher. That showed that people are *risk averse*. Bernoulli, starting from the observation that a dollar must be worth much more to a beggar than to a wealthy man, tired to explain risk aversion phenomenon by noticing that people do not value the gains in monetary values but in subjective utility, which has a property such that marginal utility is decreasing with wealth, therefore that the utility function is concave, so for instance the difference between utility of 100 and utility of 200$ is higher than the difference between 1000$ and 1100$. For this reason people value higher sure gains over some lotteries with higher expected value. By defining a concave utility function Bernoulli let TCRM overcome the St. Petersburg paradox.

Effect 2 *Risk aversion and wiggly utility function*

Friedman and Savage (Friedman and Savage, 1948) came back to the analysis of the paradox and noticed that the solution provided by Bernoulli explains only risk aversion of the type that makes people willing to buy insurance and avoid fair and unfair gambles, still does not explain the risk seeking behaviour, i.e. the preference for lotteries with lower or equal expected value to sure gains, in case of people who do gamble. Friedman and Savage noticed that a plausible explanation of the inconsistency in the form of the co-existence of risk aversion and risk-seeking would require a definition of a utility function that would have to be both concave and convex. This prompted their definition of a *wiggly* utility function which foresees that people are risk-averse for low incomes, risk-seeking for higher incomes, and again risk-averse for the highest incomes.

5.2.2 Effects captured by the Prespect Theory

Bernoulli, Friedman and Savage were constructing utility functions that covered entire wealth, i.e. a change in wealth at each choice was treated as a decrease or increase in a given amount of wealth at the starting point. An important advancement has been proposed by Kahneman and Tversky. Under their *Prospect Theory* (Kahneman and Tversky, 1979) they came up with a value function that focused on the relative

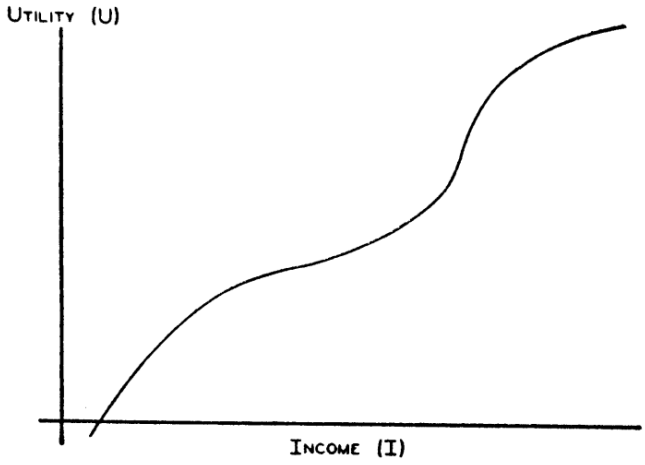

Figure 5.1: Wiggly utility function as proposed by Friedman and Savage (Friedman and Savage, 1948, p. 297)

wealth, i.e. a function that considered gains and losses as negative and positive impacts on the status quo at the stating point rather than changes in global level of wealth. This optic allows for better analysis of a single risky choice irrespective of the current level of wealth and its subjective valuations.

The prospect theory embraces several behavioural biases related to risky choices thoroughly examined by Kahneman and Tversky and reported in a series of papers. These biases are briefly recalled below.

Effect 3 *Status quo fixation*

This effect has been already outlined, as it is the basic assumption of the prospect theory. According to Kahneman and Tversky the starting point for evaluating the outcomes of events is for an agent the status quo (Kahneman and Tversky, 1984). Any gain or loss is considered by an agent as a negative or positive impact on the current state. To give a simple example if an agent losses 10$ and his starting wealth was estimated at the level of 10000$ she is not assessing this situation as becoming rich at the level of 9990$ but loosing 10$ and psychologically this is a difference that does matter. This characteristic of human decision making is likely the consequence of the qualities of consciousness. Consciousness, being based on primordial feelings and build up by both mental and biological states of brain, exists primarily in the present. It is the "here and now" that provides the background for any mental phenomena, even if the deliberation is by some means, like memory or ability to anticipate time continuum, related to the past or the future. For this reason it is natural to think of changes experienced by an agent as variations of the current state. The theory of consciousness that is accepted by many contemporary philosophers of mind assumes that consciousness can be visualized as a fluctuating plane or space that corresponds

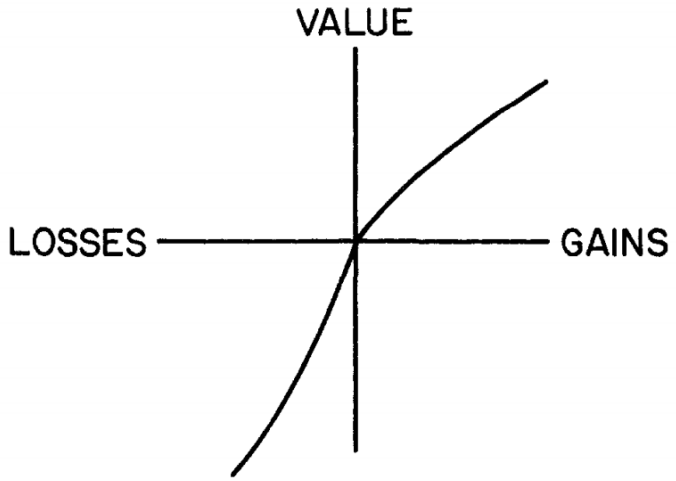

Figure 5.2: Hypothetical value function as proposed under Prospect Theory (Kahneman and Tversky, 1984, p. 2)

to the changes in the electric charge of some neurons in the brain cf. (Searle, 1999). This provides a sound background for the stance that any mental phenomena is the a change in the previous state.

Effect 4 *General loss aversion*

The value function defined by Kahneman and Tversky is concave in the domain of gains and convex in the domain of losses, however it is considerably steeper for losses than for gains which means that people in general value negatively the loss of some amount of money more than positively the gain of the same amount. Such a shape of the value function is the result of the conclusions drawn by the scientists from their empirical studies involving undergraduates who in majority would not stake 10$ on a toss of a fair coin unless the gain was not lower than 30$ (Kahneman and Tversky, 1984).

Loss aversion was one of the most profoundly studied effects (Kahneman and Tversky, 1979), (Tversky and Kahneman, 1992), (Kahneman, Knetsch, and Thaler, 1991), (Thaler et al., 1997). It has been estimated by empirical studies that that losses compared to gains are valued twice higher, to give an example the disutility of losing $100 is roughly twice as high as the utility of gaining 100$, (Tversky and Kahneman, 1992).

Effect 5 *Risk seeking in domain of losses*

The convexity of value function in the domain of losses illustrates risk seeking behaviour when an agent is confronted with choices that would result in a loss. In such a case an agent would prefer some lotteries with the expected value lower or

equal to a sure loss (Kahneman and Tversky, 1984). From the common reasoning point of view this effect embraces the heuristic that stays behind a Roman maxim *"dum spiro spero"*, i.e. "while I breathe, I hope". An agent does not what to accept a sure loss, he prefers to take considerable risk in order to avoid the loss. Having put this that way, risk-seeking in domain of losses appears psychologically consistent with general risk aversion, an agent wants to avoid losses so much that she prefers to take the risk riding on the hope for being fortunate in the lotter, rather than accepting the loss from start, even if an agent knows that in case of misfortune the loss will be higher. Risk seeking in the domain of losses has been confirmed by many empirical behavioural studies, including (Slovic, Fischhoff, and Lichtenstein, 1982), also with non monetary consequences such as pain (Eraker and Sox, 1981).

Effect 6 *Allais paradox and the risk seeking in the domain of high gains*

Allais paradox is a behavioural effect that was pinned down by Maurice Allais, who constructed a set of exemplary gambles which showed that people sometimes violate one of the main axioms of expected utility theory, the *sure-thing principle*, i.e. the independence axiom, see 5.1.2. The set includes two gambles that are presented to an agent one after another. First, an agent has to choose between getting 100mio francs (A) and a lottery (B) with three possible outcomes: 10% chance of wining 500 mio francs, 89% chance of winning 100 mio francs and a 1% chance of not wining anything. The second gamble asks to chose between two lotteries: (C) getting 100 mio with 11% probability or not getting anything with 89% probability, and lottery (D): getting 500 mio with probability of 11% or nothing with 90% probability (Allais, 1953). The expected utility for the above lotteries A, B, C, and D is $a = 100$, $b = 139$, $c = 11$, and $d = 50$ respectively. However people would not only prefer A to B, and so avoiding risk, but more importantly would violate the sure-thing principle by preferring A to B in the first gamble and D to C in the second. Obviously to be consistent with the sure-thing principle people should prefer A to B and then C to D, or B to A and consequently D to C, which was not the case in experiments carried out by Allais, which therefore show the violation of the 5th axiom proposed by Savage (Savage, 1954).

The example given by Allais falls into a more generic class of choice problems of the same structure referred to as *common consequence effect* problems. It has been empirically demonstrated that agents violate the sure-thing principle under many set-ups, regardless the outcomes are small or large, involve real or hypothetical payoffs, and the probabilities low or high (Kahneman and Tversky, 1979).

A common sense explanation of this effect is pretty simple and intuitively acceptable, as an agent seems to follow the following line of reasoning. An agent mentally visualises the situation in which she gets 100 and 500 mio francs trying to answer the question how would it be like to get 100 and 500 mio francs respectively then

estimates how likely it is for this situation to happen. Naturally she can easily distinguish between having and not having won a considerable sum of money, however estimating what is the difference in the feel-like of getting 100 and 500 is a bit harder, though it is evident that it would be judged that it is better to get 500 then 100. It would be much more difficult if the difference between these two amounts was by far smaller, e.g. a couple million francs, then probably the feel-like of having 100 and say 105 would be almost similar. The same applies to estimating the consequences of the lottery in terms of probability, it would be impossible for an agent to feel the difference between probability 0.89 and 0.9 which would be treated as equally unlikely. For this very reason when confronted with the first gamble an agent can easily imagine getting a lot of money and feels that this is at hand, she can in fact feel that she has already received and be very happy about it. Not strangely she wants to keep that feeling forever and not to take risk of losing it accepting the optional lottery. The situation is completely different in case of the second lottery. She knows right from the start that she is it is quite unlikely that she wins anything, so she is probably taking the second gamble much more relaxed and not taking is as a risk of loosing anything but opportunity to win a lot of money. And if she is to take this risk, which in both cases is pretty much the same, she would prefer to win 500 then 100 if she happens to be lucky enough to win the lottery at all. Emotions will evidently play important role in the first choice. An agent feels as if she had gotten the 100 million already, and feels happy about it, this prompts her to stick to the choice which promises the pleasure that is felt so closely. Accepting the lottery instead would mean she puts at risk something precious that her body and brain already accepted or at least mentally visualised as her own, adopting strategy consistent with general risk aversion.

The above behavioural effect has been also described and classified by Kanheman and Tversky, who present it as the second consistently observed type of risk-seeking behaviour apart from risk seeking in the domain of losses 5. According to their empirical studies people often prefer a small probability of winning a large prize over the expected value of that prospect. This effect has been also captured by the wiggly utility function by Friedman and Savage 2.

Noteworthy, Savage in (Savage, 1954) has managed to overcome the problem of TCRM with Alais paradox, i.e. the violation of the independence axiom, not by relaxing the axiom but by mere making a distinction between normative and descriptive purpose of decision analysis and emphasizing the relevance of his theoretical framework primarily for the former, by stating: "[the decision theory can first] be regarded as a prediction about the behavior of people, or animals, in decision situations. Second, it can be regarded as a logic-like criterion of consistency in decision situations. For us, the second interpretation is the only one of direct relevance, but it may be fruitful to discuss both, calling the first empirical and the second normative." (Savage, 1954, p.19). Paraphrasing these words the decision analysis as axiomatized

by Savage is primarily concerned with the question *how to make good decisions*, and the insights it provides in *how people make decisions* may be interesting but is apparently of secondary importance, at least to Savage.

So far I have been discussing effects that relate to an agent deciding about getting involved into choices under uncertainty in first place. Apart from this very fundamental decision whether to take risk at all and if so how much of it is acceptable to an agent, there are two main sources of behavioural effects under risky choices: (i) agent's subjective evaluation of the outcome, i.e. its value or utility, which is common for the riskless choices, (ii) agent's subjective assessment of probability on which the given outcome depends upon. Below I will discuss the basic effects which relate to the latter, the following section will focus on the former.

5.2.3 Effects related to subjective outcome evaluation

Effect 7 *Framing effect*

One of the most important behavioural effect from the point of view of the influence of emotion on human decisions is the framing effect. The framing effect was well described and studied by Kahneman and Tversky whose most cited paper on this topic was published in 1984 (Kahneman and Tversky, 1984). In this paper the researchers considered one of the fundamental assumption of the TCRM, namely the invariance condition. Invariance condition states that a rational agent should be able to recognize different representations (frames) of the same choice problem and should therefore have the same preference for both representations, i.e. an agent should choose in the same way regardless the framing of the choice problem. Kahneman and Tversky constructed a psychological experiment involving subjects that were supposed to judge on which of the two programmes counteracting an outbreak of a disease a state should adopt. The point was that both programmes yielded exactly the same outcomes but where frame differently, in first case the outcome was formulated in terms of numbers of lives saved and in the second in numbers of casualties. This difference made people judge significantly differently between two representations of exactly the same problem and outcomes.

There have been many similar experiments reported in the literature, of which one is particularly noteworthy, which is the so called *trolley problem* analysed by moral philosophers. It is an interesting exemplification of the framing effect because it has been procured with slightly different purpose, outside the domain of decision analysis, to study moral judgement. The trolley problem has echoed in the literature in many variants, here I will present the two classical facets of the trolley problem the *Bystander at the Switch dilemma* and the *footbridge dilemma*, as proposed by Philippa Foot and Judith Jarvis Thomson (Thomson, 1985).

Imagine a trolley coming down the track being out of control as the driver had fainted, heading towards 5 people that will certainly be killed unless the train is

directed towards the side track on which only one person is situated and would die instead. There is a switch at your hand and you as a bystander that is witnessing this situation and have clear judgement of it have to make the decision whether to change the direction of the trolley towards this single person making her die but saving the other five. The majority of people say that they would use the switch. Now imagine similar situation, yet instead of the comfort of controlling the switch you are standing on a footbridge above the truck well before the place the five prospective causalities are standing, which creates an opportunity to stop the trolley by throwing a heavy object on the track. Incidentally, there is a very fat person standing on the bridge next to you and you clearly see that you could push her on the truck and save the five people down the trucks. In this case most people would not decide to sacrifice the life of this fat person.

This phenomenon has been recently studied by neuroscientists who have discovered that these two dilemmas involve two different frames, one *personal* another one *impersonal* that result in different emotional arousal that prompts differences in the behaviour of prospects. They have found that moral judgements and decisions made in response to 'personal' moral dilemmas involved greater brain activity in the areas that are responsible for management of emotion and social cognition compared with the 'impersonal' ones (Greene, 2003). If a prospect is confronted with an anonymous, mechanistic, dehumanised so to speak situation in which she just operates a switch the consequentiality rationale prevails and the decision to use the switch is easier. However in cases where prospects imagine close physical contact with the person that is to be sacrificed a strong negative emotional response outbalances the consequentialist arguments and a prospect chooses not to do anything.

Yet again it appears plausible that agents while making real or hypothetical judgements mentally visualise the outcomes of the choice options and biologically and sensationally experience the subjective qualitative value of this outcomes, in which emotion play important role. This makes the TCRM fail to predict human behaviour in real life situations. Without involving emotional dimension into the models applied to emulating human behaviour one would have to assume the false rule that agents normally push the fat man off the footbridge and kill the baby in the basement, which contradicts the results of empirical studies.

This cases also undermine the applicability of TCRM in normative approach, as this is questionable as we see from the above discussed cases if applying the model that optimizes individual or even collective utility would be the right thing to do from the moral standpoint. Would we accept the world in which a mother would kill her baby for the sake of saving herself and a few more people? It seems that the natural answer is no.

I will come back to the trolley dilemma later in the section while discussing the emotional response to moral judgement effect 14.

Noteworthy, the already quoted neuroeconomic studies by De Martino et. all (De Martino et al., 2006) on decision making frames and biases from the perspective of the brain science confirm not only the existence of the effect itself but also its emotional background.

Effect 8 *Endowment effect*

Another implication of prospect theory closely related to the status quo bias is the endowment effect studied mostly by Kahneman and Thaler (Kahneman, Knetsch, and Thaler, 1990), according to which people value higher things they already possess to those they could have in the future. This effect may be explained by the emotion of attachement as well as searching for justification for the decisions already taken to ensure agents psychological comfort with the past decisions.

The neurological studies of the endowment effect by Knutson et.all (Knutson et al., 2008) suggest that it is primarily driven by loss aversion regarding the owned or preferred products, therefore it is the anticipated salience of loosing the already possessed or preferred product that causes the effect, which again takes us back to the emotions contributing to loss aversion effect.

5.2.4 Effects stemming from subjective assessment of risk

Effect 9 *Headlines effect*

As this heuristic has no common name in the literature I will refer to it as the "headlines effect". This heuristic refers again to the assessment of probability of events, which is that people tend to overestimate the probability of events when they take excessive account of information that is salient, easily available or to which extensive attention by mass media is paid to. The most common example of this heuristic is: people tend to perceive flying by plane more risky than driving a car because non-occurrences of flight accidents are not reported meanwhile occurrences are widely communicated over the mass media.

The headline effect is the good exemplification of how experience and emotion bias the individual choice. People rely on experience, they judge probability based on the events they experiences not based on the abstract analysis on the problem space. Furthermore they assign high emotional negative value to the occurrences of plane crashes as these are typically covered by all news channels reporting high number of casualties and low survival rate of passengers. The emotional load of this type of news is fuelled, often purposefully, by detailed news coverage, breaking news mode, and drastic shots form the plain crash scene.

Effect 10 *Source dependence*

The *source dependence* effect describes a phenomenon that agent's subjective assessment of the probability of an uncertain event depends not only on the degree of the objective probability but also on how this probability is derived, what is its source. In an experiment procured by Ellsberg (Ellsberg, 1961), referred to as *the two-color problem*, people where asked to bet on urns that contained 100 black and red balls. It was revealed that urn A contained blacks and reds in a ratio unknown to anyone meanwhile urn B contained and even split of balls of each colour. It turned out that prospect would bet on the second urn more often, assessing the chances to get a desired colour of the ball from the second urn regardless the colour, therefore relying on or 'trusting' urn B more regardless the gamble offered based on these two urns. This at the same time constituted a counterexample for the TCRM's axiom of subjective probability additivity to 1.

This effect, also known as ambiguity avoidance , was later studied by Heath and Tversky (Heath and Tversky, 1991) who showed that the effect also occurs in situations when the source of ambiguity lies in the prospect's knowledge. More explicitly, they conducted a series of experiments in which they let prospects choose between clear chance events of unambiguous probability and events depending on their beliefs, which showed that people favour clear chance over uncertain beliefs if they feel incompetent in the domain of the choice problem, contrary to the situation in which they feel competent and would rely on their beliefs instead.

Clearly ambiguity relates to the emotion of fear, the basic emotion responsible for survival (Darwin). Fear is one of the basic emotions and the fear of the unknown one of the five basic fear factors (Ollendick, 1983). The ambiguity avoidance likely has the emotional background. Fear triggers basic bodily emotional responses such as 'freeze' or 'run', is definitely a negative emotion that an agent wants to avoid. This well explains the effect of ambiguity avoidance, as the known would be always preferred to the unknown, and as it is strongly emotionally conditioned the rational argument that the objective probability of choosing from either urn under the Ellsbeg's two-color problem is equal would be largely ignored by prospects and urn A would be avoided. Evidently, fear of the unknown also provides a good justification of risk avoidance at large as involvement in a risky decision implies accepting an unknown outcome of the choice.

Fox and Tversky (Fox and Tversky, 1995) showed that ambiguity avoidance is much stronger when an agent compares a clear and ambiguous option under one choice compared to the situation when an ambiguous chance is assessed individually. This well corresponds to the way people deal with emotion of fear, once it cannot be avoided an agent would face the situation and look for a best response to the situation. Consequently emotional reaction in choices under ambiguity would bias the decision to a greater degree in situations when ambiguity can be avoided.

Insights from recent neurological studies, involving fMRI brain activity monitoring of prospect participating in behavioural experiments, indicate that choices under

uncertainty involve increased activity in orbitofrontal cortex and amygdala, areas of the brain responsible for emotional and motivational processes, which confirms that emotions do play important role in decision making. Although according to some studies it is not evident that risk aversion always has emotional background (Tom et al., 2007), it is evident that some types of risk-aversion, in particular that observed under ambiguity more then under risk, prompts increased activity of amygdala responsible for emotional orchestration in the brain, which supports the claim that fear underpins ambiguity avoidance (De Martino et al., 2006).

Effect 11 *Naive diversification heuristic*

This hehavioural effect is discussed by Bernatzi and Thaler (Benartzi and Thaler, 2007) primarily in the context of choices about pension schemes. A couple of studies, incl. that involving UCLA proffersors deciding about the composition of pension fund portfolio reveal that the initial composition of the offered menu of pension largely impacts the eventual proportion between the chosen saving instruments. People while allocating savings tend to split equally between the offered compositions of saving instruments. For instance if they are asked to allocate 100 euro and are offered with a choice of 2 instruments they would tend to split equally between them 50/50, if they are offered 4 than the decision would be biased towards 25/25/25/25 split. The paper recalls a justification of a savings allocation decision by Nobel laureate Harry Markowitz, a founder of modern portfolio theory who once admitted:

> "I should have computed the historical covariance of the asset classes and drawn an efficient frontier. Instead I visualized my grief if the stock market went way up and I wasn't in it - or if it went way down and I was completely in it. My intention was to minimize my future regret, so I split my [pension scheme] contributions 50/50 between bonds and equities." (Zweig, 1998)

They also quote and discuss other studies including that by Read and Loewenstein (Read and Loewenstein, 1995) involving children, trick-or-treaters during Halloween, which evidenced strong diversification bias in situations where children where offered to chose two candy bars from two types at one time (in the simultaneous choice condition), compared to the situation when they were offered to chose one bar at the time at two adjacent houses (in the sequential choice condition). The justification of the decision on pension scheme contributions by Markowitz suggests that naive diversification heuristic is a type of loss aversion. The regret of incurring loss in case an outcome alternative to the expected one turns out to be true makes prospects bet on a 50/50 gambles. In a way betting on a 50/50 chance gambles may be treated as avoiding entering into risky choice at all by simply refraining from subjective assessment of probability, relying entirely on faith in this respect. This may be true

for an agent that wishes to avoid a situation in which one would have to blame oneself for a wrong decision stemming form miscalculation of risk.

Effect 12 *Myopic loss aversion*

This effect has been dubbed by Benartizi and Thaler (Benartzi and Thaler, 1995) as an explanation to the equity premium puzzle, later empirically validated by Thaler et al. (Thaler et al., 1997). Myopic loss aversion is am effect closely related to loss aversion and refers to the sensitivity to losses combined with a tendency to evaluate outcomes frequently and adjust decisions accordingly with the focus on shorter time perspectives. This effect was observed in the analysis of pension scheme decisions, where prospects had a chance to revise their investment portfolio during the investment period. I was observed that investors tend to take loss averse decisions each time they receive information on the investment results so that the frequent feedback provision led to lowest risk portfolios and ultimately poorest return on investment. The study of this effect is in particular interesting because it proves that experience directly impacts the decisions. The utilities and objectives of the investors remained the same over the period still the sole fact of getting new information and opportunity to reconsider decisions made strongly influenced the end result.

Effect 13 *Major decision effect*

Benartzi and Thaler demonstrated in their studies of pension saving behaviour (Benartzi and Thaler, 1995; Benartzi and Thaler, 1999; Benartzi and Thaler, 2007; Thaler et al., 1997) that people tend to apply heuristics to important decision. This can be observed based o the amount of time taken by prospects deciding on the pension fund portfolio definition. In one of the above-cited studies by Benartzi and Thaler (Benartzi and Thaler, 1999) professors were observed to take such decisions in less than an hour on average.

Major decision effect presents a deviation from TCRM, which assumes agents are pervasive utility maximizers. Many empirical behavioural studies show that when people are confronted with a "major" or important decision they tend to behave in a way that cannot be explained in a rational way applying all sorts of heuristics, including imitation and naive diversification. Important decisions are those that imply intensive emotional states for which reason agents tend to apply heuristics instead of performing a thorough normative analysis. Ambiguity aversion would also contribute to the explanations of this effect.

5.2.5 Effects considered in philosophy of mind and moral judgements

Although detailed discussion on the practical reason and philosophical perspective on agency has been already discussed in detail in chapter 3, for the purpose of

consistency and completeness main effects identified at those grounds are presented below, together if some that appear for the fist time.

Effect 14 *Emotional response in moral judgements.*

Greene at al while carrying out studies on neurological underpinnings of moral judgements proposed an example of a moral dilemma similar to the *trolley dilemma* 5.2.3 that is referred to as the *crying baby dillema* (Greene et al., 2004). Imagine that enemy soldiers have taken over your village and they are ordered, which you have heard, to kill all remaining civilians. You have hidden with some other people in a cellar. Soon you are hearing the soldiers approaching as your baby has just started to cry loudly. You cover her mouth with a piece of cloth and you have to decide either to uncover her mouth to let her breathe which would alert the soldiers and inevitably cause them kill everyone including you and your child, or smother your child to death. The study of brain activity of prospects confronted with such as dramatic moral dilemma let Greene and his colleagues conclude that while making moral judgements there are two parallel mental processes activated in drive people away form personally harmful actions, whilst the second involves the 'cognitive' part of the brain associated with higher cognitive functions: executive control, planing, reasoning and making economic decisions. These two opposite psychological processes compete so to speak, and depending on how much conseqeuntialist rationale versus emotional response is there the corresponding part outweighs and prompts the final choice.

Effect 15 *Cynical decision strategy*

Sloterdijk (Sloterdijk, 1987) introduced a contemporary archetype of *homo cynicus*, i.e. a representative of contemporary culture where cynicism is the dominant mode, a person characterised by the "enlightened false consciousness". It is likely that a homo cynicus would internally assume certain true preference ordering. However the axiom of TCMR on following this ordering in choice situations would likely be violated as homo cynicus could purposively or spontaneously choose options not in line with the preference ordering. The same could apply to choices in strategic, competitive situations as considered by the theory of games (Von Neumann and Morgenstern, 1944), when a cynical agent could play a dominated strategy to mislead, provoke or astonish other players who expect each other to behave rationally, so that the common knowledge and equilibria could not be easily predicted.

Effect 16 *Acting upon total reason for action*

Searle (Searle, 2002) presenting his critique of TCRM introduced an alternative decision model which he grounded on the theory of intentionality and speech acts.

According to Searle instead of maximising utility a human agent, more precisely the *irreducible self* of an agent, makes decisions based on the reasons for actions that are intentional states of motivational character. Importantly there is no reason such that it could provide causally sufficient condition for consciously voluntary action, i.e. there is always a gap of free will (cf. next effect). Putting the model in simple terms (for details please see chapter 1) an agent in a deliberative process considers: valid reasons for actions (incl. trade-offs between conflicting reasons), effectors and constitutors (elsewhere in the literature this process is referred to as means-to-ends analysis), and makes choices in such a way as to satisfy/not compromise as many of these reasons as possible without compromising other reasons, which are either desires or commitments.

Saltzman and Newsome (Salzman and Newsome, 1994) research on neural mechanisms for forming perceptual decision shows that a deciding brain presents increased neuronal activity in certain of its parts as if it 'accumulated' simultaneously arguments for different available options, the choice is made the moment one options 'prevails' which is manifested by visibly strongest neuronal activity in one of the parts taking part in this 'neuronal dispute'. This suggest that decisions are made based on conscious evaluation of available options by internal collection of arguments for and against available alternative, which corresponds well to the theoretical propositions by Searle.

If this is indeed how people take decision then the deliberation, which takes place in the unified field of consciousness, would be naturally impacted by related aspects of consciousness, including emotional states. This puts the decision making into a significantly different, very dynamic, set-up compared to TCRM, which relies on subjective utilities that are stable, consistent also in time.

Effect 17 *The gap of free will*

Finally, one of the main conclusions from the discussions in Chapter 3 on the free will is that at the level of analysis adopted in this section the gap of free will, regardless one perceives it metaphysically as an illusion or not, constitutes and important, if not the most important behavioural effect. The consequence of this effect for decision theory is indeed dramatic, as it implies impossibility of deductive theory of action, i.e. no fully robust algorithm for human action can ever be defined (cf. (Searle, 2001)). This means that however accurate a model of human behaviour is there will never be certainty about its predictions. However as Searle puts it "The causal gap does not imply explanatory gap", so although we can say what where the reasons for actions we cannot derive actions from reasons. Still, it seems plausible to assume that before the gap of free manifests itself in the form of counter rational behaviour the premises of rationality (reasons for actions) will be taking effect, consequently the more we know about what constitutes rationality, including

how conscious experience with its emotional background influences rational choice the better the models we can construct.

The above presented overview of behavioural effects, for the limited space merely providing some examples rather than a complete catalogue, shows that these effects could be better understood and explained by linking decisions with affective and experiential phenomena of a conscious human being. The TCRM cannot resist the bulk of evidence against it and new models of agents rationality must be sought for.

5.2.6 TCRM in ashes

Simon and Newell (Simon and Newell, 1958; Simon, 1973) have told apart the two main classes of decision problems: *well structured* or *well defined* problems from *ill structured* or ill defined problems, of which the former, unlike the latter, comply with the classical modelling assumptions. For a couple of past decades the efforts of decision analysts who where confronted with ill defined problems have been concentrated on developing appropriate methods for structuring the ill defined decision problems, so as to translate them into well defined problems and be able to apply mathematical analysis apt for TCRM. Noteworthy, it is a common strategy taken by normative decision theorists, who continue exploiting the TCRM or its more recent variations, to work with the problem space and the preference function in particular so that it addressed the ill structure of the real problems better. Often for the sake of mathematical elegance implausible assumptions are accepted, such as those implied by the TCRM, with little or no justification. Meanwhile as it has been provided across this thesis there is enough psychological, philosophical and neurological evidence to falsify TCRM on the positive, descriptive ground. Natural agents simply do not take decisions as TCRM assumes, so it is not informative enough for if one wants to emulate natural agent behaviour in an artificial system. Though it could of course be taken under normative economics as an axiom telling how an artificial agent *should* behave to be rational in the TCRM sense.

Still, it must be admitted that the long years of TCRM's dominance in economic and AI thinking is due to the lack of alternative models that would fill in the gap. Persky rightly notices:

> "...to compete successfully against economic man, a new ethology must be parsimonious; it must clearly specify the relevant psychological make-up of economic agents; and it must demonstrate that such a system yields better and/or new insights." (Persky, 1995)

Undoubtedly, there is a need for a model of agent rationality that would embrace agent's affective states and qualities of conscious experience, as clearly affect is the basis for meaning, which was a conclusion from chapter 3 and catalysts for human behaviour.

Affect manifest itself in the decision making process in two ways: (i) as remembered affective value of experiences and (ii) as the affective reaction to current stimuli. Separating these two effects on behaviour is difficult given the homogeneity of consciousness, but distinction is possible given the analysis done at the level of neuronal systems in the brain, which could provide sound background for a new theoretical model of behaviour. The remaining part of this chapter will be dedicated to alternative, contemporary models of rationality, with particular emphasis on the *BDI* rational agency model and will reflect on how the experience representation framework proposed in chapter 4 could be used to improve the existing approaches to modelling emotional agents and prompt the formulation of a rationality model of an experiencing agent.

5.3 Contemporary approaches to modelling rational behaviour

Currently we can distinguish two main approaches in applied studies of human decision making: *probabilistic-behavioural* approach and *neuro-cogno-logical* approach. The former perceives a rational agent as a 'black box', ignores entirely its inner processes, reckoning these as to complex to understand and model, only focusing on what comes in and out of this 'box', hence what are the inputs (external environmental states, conditioning variables) and outputs implemented decisions (actions). This approach involves modelling decision process by applying games theory. Each agent has a set of available choice options each to be chosen by the agent with a given probability. The probability distribution is assessed based on observation, which is where behavioural analysis comes in. With modern information, computer and digital network technologies the abundance of user data and excess of computation power makes this approach very attractive as allows to make more and more sophisticated algorithms estimating decision options available to the agent as well as probabilities of an agent to act upon each of them. This model has both advantages and disadvantages. The main advantage is that, given the calibration of model is done based on sufficient amount of relevant data, the accuracy of models can be surprisingly high, which is why they can be applied in practice. However an important limitation of this approach is that it can hardly say much about any anomalies, abnormal behaviours, especially in situations that are not standard. Furthermore they work much better on macro scale than on micro scale, as the largest the group of modelled agents and the longer the time horizon the higher the accuracy. Individual choices are much more difficult to predict but with scale the models start to work better.

The other approach confronts the challenge from the problem solving perspective. It primarily addresses the question "How people choose and act upon their choices?". By comprehensive understanding of the reasons for actions it is believed to be able to

come up with better algorithms predicting actions taken by agents. The algorithms in question would emulate agents rationality and thus make satisfactorily accurate forecasts on their actions. Here the key techniques used for modelling are delivered by machine learning theory, agent-based systems, formal logic and linguistics.

Importantly, in principle there are no unbridgeable gaps preventing from unification of these two approaches under frameworks that emulate both behaviour in competitive social situations and individual decisions, however the two opposing schools show little confidence in the findings but primarily theoretical underpinnings of other approach.

To complete the picture it should be mentioned that some proponents of cognitive approach, in particular some more orthodox neuroeconomists suggest to dive even deeper into the neurological processes behind human choice, as the key to the human behaviour may be found at the level of synaptic processes, or at the higher level of brain regions or systems of regions that normally are looked into in functional studies of the brain. This provokes major sceptical criticism from mainstream economists who doubt that neurological insights could significantly improve economic normative models and pave thus a way to more efficient practical applications. One of the main arguments raised in this respect is that the opening of "the black box" is an endless exercise because there is not a single box but rather a Russian doll, so that models of neural processes are black boxes of lover cellular processes, atomic processes, subatomic processes and so on, of which little sense can be made at all, (Bernheim, 2008). Importantly however, it is beyond doubt that neuroeconomics provides valid empirical clues for constructing economic and social theories, and it has been so ever since cognitive and brain science emerged.

5.3.1 Unification of behavioural sciences under contemporary game theory

There two important exemplifications of the above identified families of approaches to modern rationality modelling. In this subsection we will present the account by Herbert Gintis as outlined in (Gintis, 2009).

Gintis, who builds on the TCRM tradition, has realised that the issue of economists failing to model consumer behaviour lies not in the faults in the Bayesian rationality model itself, but primarily in ignoring the social context of individual choices. Bayesian rationality has served as the basis for analysing rational agents behaviour in competitive situations under games theory since the seminal work by Von Neumann and Morgenstern (Von Neumann and Morgenstern, 1944), yet it has been entirely ignored that rational agents may share mental constructs. So Gintis emphasized that the assumption that humans are rational is a solid fundamental approximation but human agents function and take decisions in a social context, they are not bounded by their subjectivity. Here is how he puts it:

> "Humans have a social *epistemology*, meaning that we have reasoning processes that afford us forms of knowledge and understanding, especially the understanding and sharing of the content of other minds, that are unavailable to merely rational creatures. This social epistemology characterizes our species. The bounds of reason are thus not the irrational, but the social"

This allows Gintis to come up with a handful of patches for TCRM and thus explain many behavioural effects by taping into social psychology. Importantly, Gintis sees social reality, including social norms, as irreducible phenomenon that cannot be derived from a model of interacting rational agents, not mentioning the lower level processes inside individual decision maker's mind. This is accompanied under his account by the believe that the evolutionary approach to strategic, competitive interaction can fully explain rational behaviour. So Gintis suggests that evolutionary game theory (Smith and Price, 1973) may provide the unification framework for studying natural agents behaviour across behavioural sciences. Rational agents under such a framework could for instance copy strategies that ensure highest fitness, these strategies would thus diffuse across populations of players instead of being deduced by individual rational agents, especially when such deduction (information processing) incurs high costs.

So the framework which Gintis proposes is based on five key elements: (i) gene-culture co-evolution; (ii) the sociopsychological theory of norms; (iii) game theory; (iv) the rational actor model (modern version of TCRM assuming Bayesian rationality); and (v) complexity theory.

Importantly, Gintis rests on Bayesian rationality as fundamental principle for starting analysing rational behaviour, which is bounded by society and culture. He emphasizes that diversion of behavioural scientists from Bayesian rationality model has led to "theoretical disarray", resulted in no alternative improved frameworks and thus provided no new useful insights.

The position by Gintis in so far as TCRM is concerned, which is typical of proponents of Bayesian rationality model, is well illustrated in juxtaposition with the views presented by followers of Simonian, psychological approach. For instance to Krantz assumption:

> "The normative assumption that individuals should maximize some quantity may be wrong. ...People do and should act as problem slovers, not maximizers." (Krantz, 1991),

Gintis replies:

> "This is incorrect. In fact, as long as individuals are involved in routine choice and hence have consistent preferences, they can be modelled as

> maximizing an objective function subject to constraints." (Gintis, 2009, p. 236)

At the same time the seeming flaws of the model can be dissolved with incorporating social and evolutionary context into it. He for instance underlines that some behavioural effects uncovered by Tversky and Kahnemann stem from the fact that people do not simply know how to calculate objective probabilities in risky lotteries so they are guided by their common sense instead, once however instructed how to calculate objective probabilities agents would follow the choices dictated by mathematically calculated expected value.

> "If humans fail to behave as prescribed by decision theory, we need not conclude that they are irrational. In fact, they may simply be ignorant or misinformed." (Gintis, 2009, p. 7)

This statement by Gintis is pivotal. This statement implies that not behaving according to utility maximization is wrong, detrimental to well-being of conscious creatures. It implies that there is a common recipe for "proper", "normal" or "fit" behaviour which is that obeying the prescriptions of decision theory. This could hold only if decision theory could come up with a recipe that embraces all human behaviour. The key here is that often decision theorists presuppose that there must be a universal rule of behaviour, in case of Gintis these are the rules of Darwinian natural selection captured by evolutionary games theory. This is practical, but unlikely to be true. It requires moral and subjective judgement to state which way of human conduct is normal, i.e. which rules of conduct are acceptable from the point of view of a "rational" agent. We could easily try to think of some rules that hold for the majority of people such as: people prefer to live than to die, people prefer more than less goods, people pursue happiness, but if we consider the case of a suicide-bomber, a suicide in despair, a Buddhist monk, a masochist, we would have to rest on stating after Gintis that they commit as so called "performance error" or are "misinformed" or "ignorant", which is not convincing. Of course, it could be said that a suicide-bomber optimises social utility sacrificing his life for the well-being of his community, but even if we are able to come up with rules of such low granularity at the level of individual agents would the tools of decision theory be of any use? No, because they only make sense if there is at least a group of people obeying the same rule if we want to explain these cases with social norms.

This is not to undermine the utility theory at large, as indeed we can define a large space of problems that can be neatly addressed with it, however we must remember that relying on the axiom of preference consistency has its price, which is the elimination of behaviours that do not fit this axiom from the scope of analysis, regardless how we try to tweak the problem space as to make it more and more inclusive.

Nonetheless, the method proposed by Gintis to work with the problem space in a way that we are able to define rules, preference consistency in particular, that hold for more choice situations seems the valid one to follow. An example is apparent time-inconsistency of preference. If we modify the choice space to include payoffs at different points in time the inconsistency in preference will disappear. We propose that if we include experience in the scope, i.e. if we consider affective accompaniment to knowledge, a few more inconsistencies could be eliminated. The emotional background of conscious self composed of primordial feelings represented in the upper brain stem as suggested by Damasio (Damasio, 2010) supports the stance that feelings play fundamental role in any conscious process.

Gintis says that indeed anomalies occur but if we take this as general rule this would mean that we believe most of people are psychopaths. It is unlikely however that the claim that most or many of us are psychopaths or at least that there is important amount of psychopathic behaviour such that it influences socio-economic processes to the extent that it influences the emergence of unpredictable phenomena that impact the majority of us should be considered false. All depends how we define what is normal.

Gintis after Robson (Robson, 1996) argues that preference consistency flows from evolutionary biology, citing research results on nonhuman species, including insects and plants that show applicability of decision theory to their behaviour. In evolutionary biology the behaviour of individual organisms is guided by preserving the fitness of organism, which is related to its expected number of offspring. Gintis, going back to Darwin (Darwin, 1872) highlights that organisms do not directly maximize fitness but have preference orderings that are themselves subject to selection according to their ability to promote fitness. This seems an idea that is consistent with the theory that says that human brain as an organ of the body, has been created in course of evolution to give birth to consciousness and free will which underpin the most efficient (from the evolutionary perspective) decision apparatus responsible for promoting fitness. However, there are two principal problems with this approach. First, minor one is that we still cannot be sure that this is the selection of preference orderings that govern evolutionary adaptation. The second, far more important, is that even if in the long term the above claim holds, in the short term we have to include in the behavioural analysis all the organisms, also those that are not fit and will become extinct in one or more generation. This is very important because we want to address with decision theory the problems that we are facing now, not solving problems in the hypothetical future state of the universe where all organisms are fit. Needless to say there will probably never be such a point in time where there is no place for fitness improvement, as apparently the evolution process is an endless one.

In conclusion, approach presented by contemporary proponents of TCRM such as Gintis appears promising it so far as it enlarges the scope of Bayesian rationality

beyond a self-regard individual to a social game, still the approach of evolutionary game theory operates with an idealized model of rational choice that cannot embrace all human situations.

It is important to highlight that Gintis conclusions are in line with our earlier observation that there is a potential for synergy between the Bayesian rational actor model and deliberating problem-solver model favoured by psychologists. He notices that this synergy could appear when the TCRM model is applied to routine choice and the problem solver approach to more complex human deliberative decisions, goal formation and learning (Gintis, 2009, p. 238). Furthermore Gintis insists that to make important advancement in economic and behavioural theories scientists must adopt agent-based models, and theoretical model verification against empirical data.

In the next subsection we will consider contemporary frameworks for rational agency representation in information and computer systems underpinned by the problem-solver model, which has emerged from the pioneering work by Simon and Newell (Simon and Newell, 1958; Simon, 1959; Simon, 1978; Newell and Simon, 1961).

5.3.2 Deliberating agents

> "Likewise, to predict the short-run behavior of an adaptive organism, or its behavior in a complex and rapidly changing environment, it is not enough to know its goals. We must know also a great deal about its internal structure and particularly its mechanisms of adaptation." (Simon, 1959, p. 255)

The idea that behaviour of an intelligent natural agent, seen as problem-solver, which uses its mind's information processing capacities to manage life could be emulated in a computational formal system was pioneered by Simon and Newel (Newell and Simon, 1961). In the times where computer functionalism started to take over it was tempting to draw this parallel, especially as TCRM, then in the mainstream, was going through difficult times confronting more and more counter evidence being fostered by behavioural research. If mind is to brain as a programme to a computer and human thinking is information processing, why not construct a computer programme like human mind that could solve problems in an equally intelligent manner? Although it has taken years to prove it is a more challenging task than assumed initially, which to this date has not been solved, the parallel turned into the AI vision and a Holy Grail of the field, and the goal of constructing an artificial rational agent is still being pursued.

The problem-solver approach to human decision making starts with an assumption that the key to understanding human behaviour is to understand human mind. It is important to find out how rational agents reason about ways for achieving goals and what motivators determine the selection and pursuit of goals. In other words what is

the internal mental structure of a rational agent, and what makes up for rational agency in first place.

In contemporary information systems and computer science, AI in particular, agent-based approach is the mainstream method for representing human practical rationality. This method involves agents, rational or intelligent agents, that are a certain class of computer systems. These systems are called agents precisely because they have some features typical of a rational human decision-maker. So far when we have been talking about agents the epistemic understanding of the term should have been applied. From now on the term agent may appear in a new meaning: an artificial system, precisely a computer system, that represents behaviour of a human rational agent. To avoid confusion we will use the term *artificial agent* (AA) for this new meaning of agent.

Artificial agents have the following main features: (1) they make decisions, therefore they act, so they can have causal effect on the environment; (2) are autonomous in their behaviour and (3) are rational, i.e. they follow certain rules while acting, in other words they make adequate not random decisions, furthermore typically artificial agents (4) are capable of communicating with external world and other agents, so they sense and influence (simply act, as sent communicates can be seen as speech acts) (Wooldridge, 2000), (Russell and Norvig, 2009). Wooldridge and Jennings add one more important feature: proactiveness, which apart from covering the core feature of performing actions, embraces also the exploitation of serendipity, i.e. being able to grasp unexpected opportunities to initiate fit actions (Wooldridge and Jennings, 1995). Artificial agents can be seen as complex input-output process or a box that processess inputs to outputs, however for problem-solver approach what is in the box is of primary importance.

The basic idea about modelling human behaviour with artificial agents is relatively simple. As humans can be defined as rational decision makers behaving according to a set of observable rules it should be possible to construct an artificial agent that would have a particular set of rules encoded together with some metarules setting out how new rules and knowledge is acquired by an agent, so that the given artificial agent could represent the behaviour of a human being. Importantly, as agents can interact with the environment, other agents in particular, complex socio-technical systems can be modelled and social behaviour emulated by applying agent based systems, which is referred to as agent-based modelling (ABM). Moreover, AAs can learn, taking advantage of machine learning algorithms, and therefore evolve along the way, altering behaviour patterns based on the updated knowledge. Repetitive interactions between agents allow an agent-based system exhibit complex behaviour patterns and lead to emergent phenomena (Bonabeau, 2002).

Consequently, the key questions under the considered account are: what is the agent process like?, what is that drives the behaviour of an artificial agent?, how to define the rules of its behaviour? In answering these questions information AI

builds on knowledge regarding rational agency from many disciplines, primarily from philosophy of mind, behavioural psychology, decision theory, neuroscience and formal logic.

Importantly, in computer science an artificial agent is a computer programme accompanied with some architecture (Russell and Norvig, 2009). An agent program implements the agent function which is the mapping from percepts to actions. The architecture includes the computing device and other hardware such as physical sensors and actuators. [1]

Taxonomy proposed by Weiss (Weiss, 2000) distinquishes four main classes of agents: (1) logic-based agents, (2) reactive agents, (3) belief-desire-intention agents, and (4) layered architectures. This taxonomy provides a useful distinction between layered architectures that represent hybrid approach combining different realizations of AA's decision making via various software layers from all other classes. Distinctions between reactive agents, logic-based agents and BDI agents are somewhat more problematic, as these classes seem conjunctive. However it provides a good overview of what are the possibilities of architectural implementations of AA rationality. All in one there are three possibilities: (1) reasoning based on formal logic; (2) optimisation of an choice function, however in a reactive relation to environment (otherwise a regular transformational system not agent-based system); (3) hybrid/multi-layer approach. So, in all the cases AA population is modelled as a mulit-agent system where each natural agent is represented by one AA. In first case a formal system is introduced to represent the way how an AA reasons about actions. This should be based on a plausible theory of customer behaviour. Therefore a formal language, axioms and a set of inference rules must be defined. Typically for this purpose existing formalisms are adapted and reused from within either monotonic and nonmonotonic logic or new formal systems are defined. Under this class the way how a given agent behaves is translated into a set of axioms and theorems, valid formal utterances, which constitute the basis for deducting the way the customer is going to behave in a confronted situation. With regard to the second possibility, a choice/action function is defined. Typically an agent would have a SEU function defined so that depending on the inputs from the environment (constraints) AA choices can be deduced under the supposition that AA optimises subjective expected utility. The major challenge is to define this function in a way that it describes well the behaviour

[1] It is worth underlining that any AA, from the software engineering point of view, is a reactive system, as opposed to a transformational system. The distinction made by Harell and Pnueli says that a transformational system is a system that in principle takes some input and processes it by a given function by which it produces outputs after which it terminates. On the contrary, a reactive system intensively interacts with its environment, it repeatedly monitors the outside world and continuously responds to the external inputs, so a reactive system instead of performing a function "maintains a certain ongoing relationship, so to speak, with its environment" (Harel and Pnueli, 1985). Such a description of a reactive system well corresponds to the nature of a human agent, and well addresses the intrinsic dynamism of real-world problems, on which traditional static methods of decision theory concentrated on optimization of a utility function fail hopelessly.

of AA. Finally a mix of both approaches could be implemented by means of a layered architecture. An additional challenge in this final setting is to orchestrate communication between different layers as well as to come up with a plausible agent behaviour theory that would fit the way these layers are interconnected and provide for meaningful interpretation of results.

Wooldridge (Wooldridge, 2000) proposes that to be able to define a rational AA we need three fundamental components: (1) a plausible theory of rational human action (human rational agent); (2) a computer system implementation of an AA; (3) a logical component that allows to reason about AA's behaviour formally, where the main challenge is to axiomatize the properties of agents. Components (1) and (3) are particularly relevant for the thesis, meanwhile (2) falls evidently out of its scope. In the following subsection I will review (3) how modern AI addresses the formal representation of AAs rationality, insofar as the logical component is concerned, which appears as the main challenge for AA architects. Later I will concentrate on AA models that take affect into account to find out if the available frameworks are sufficient for implementing the model of a rational *experiencing* agent. For the remaining part of this subsection I will concentrate on (1) presenting mainstream model of rational agency. AA architects visibly concentrate efforts on AA implementations that is components (2) and (3), meanwhile to us these are the arguable theories of human rationality that are the weakest points in the rational behaviour emulation systems.

The mainstream model of rational agency currently is the model of deliberating agents, the so called *believes, desires, intentions (BDI)* agents. The model is supported by a practical reason theory proposed by a philosopher Michael Bratman known as the planning theory of intention and agency (Bratman, 1987). Bratman's account starts with the following assumption:

> "We are planning agents. Our purposive activity is typically embedded in multiple, interwoven quilts of partial, future-directed plans of action. We settle in advance on such plans of action, fill them in, adjust them, and follow through with them as time goes by. We thereby support complex forms of organization in our own, temporally extended lives and in our interactions with others; and we do this in ways that are sensitive to the limits on our cognitive resources." (Bratman, 1999, p. 1)

These reminds us of conclusions from classical works in practical reason and purposive human action, like that by von Mises discussed at the beginning of this chapter, however it already incorporates elements of computer functionalism, which is visible in approaching rational agents as information processing systems producing action on the output side. It is important to mention that the Bratman's theory like most of contemporary philosophical accounts is strongly influenced by the intentional stance, discussed in chapter 3. The consequence of adapting the intentional stance is

the acceptance that rational agents have mind states that are somehow related to or are about external world. [2]

As Hoek and Wooldridge notice rational agents can have all sorts of possible mental states such as: beliefs, goals, desires, intentions, commitments, fears, hopes, and obligations among others, and propose to group them into three clusters of attitudes:(i) *information attitudes*, including mostly believes, (ii) *pro attitudes* that invoke agent's actions so primarily desires, intentions and goals, (iii) *normative attitudes* including obligations, permissions and authorization (Hoek and Wooldridge, 2003). Under Searlean account these three would correspond to (i) believes, (ii) prior-intentions and intentions-in-action, and (iii) reasons for actions comprising desires and desire-independent reasons for actions: commitments and obligations (Searle, 2001). In broad terms human and nonhuman agents are considered rational purposive agents who pursue goals given their mental representation of the world, their constantly figure out which goals (ends) they want to pursue and combine means to ends in the process of deliberation and means-to-ends reasoning, and thus coordinate action. Action is therefore motivated by desires and other non-desire reasons for action, but desires and other reasons for action can be inconsistent and the agent may not know by which means these could be satisfied (Weiss, 2000, p. 344). So an agent first of all has to figure out which of the potentially conflicting reasons for action to accept, which is about selecting those that are consistent and achievable. This process is known as deliberation which leads to selection of *goals*. Goals provide basis for *intentions* that are planned actions. Importantly Bratman underlines that the capability to engage in planning, i.e. future oriented intentions is which distinguishes human agents from other purposive agents. Importantly, planing comes in stages, or there are different levels of plans, as agents are flexibly confined by these plans, not all of them are turned into action, which is where intentions and prior intentions and finally intentions in action (as Searle puts it) comes in.

The distinction between goals chosen for pursuit and all the other goals is very particular for BDI framework which emphasizes in practical reason the significant role of *intentions*, which are simply goals to which an agent committed itself. Bratman does not distinguish goals as separate entity from intentions but rather emphasizes there are different types of intentions, which depend on how and for how long an agents commits to it. *Commitment* therefore decides on the *persistence* of an agent in pursuing intentions (goals) and putting them in action. [3] Goal remains however a

[2]Especially here it is important not to confuse *intentionality* of mind with a common meaning of *intending*, which relates to action, so planing to do something. Intentionality of mind means the unique capacity of mind to be about external world, in other words the capacity of mind to have mental representations of external reality. For detailed explanation please refer to chapter 2.

[3]Noteworthy, the commitment we are talking here is an internal, subjective attitude of an agent, and should not be confused with commitments understood as externally directed intentional states that constitute a desire-independent reason for action, e.g. agent A has committed to read a book tonight.

useful technical term in BDI agents implementation frameworks used for separation of particular type of intentions. It is important to highlight after Wooldridge that intentions are *future-directed* states of mind, not actions themselves (Wooldridge, 2000, p. 23). Searle in turn highlights the role of a special kind of intentions that allow an action to be continued and completed, the so called *intentions-in-action*. And so we could distinguish: goals, prior intentions or partial plans, and intentions-in-actions.

For this general introduction on BDI framework it is enough to take away that agents operate with three basic mental states: believes, desires (and after Searle desire independent reasons for actions such as commitments or obligations), and intentions which are future oriented plans of actions or mental states that allow an agent to persist in action (intentions-in-action). The persistence in intending is dependent on internally directed commitment of an agent to pursue an intention. It is important to highlight that an agent manipulates all these mental states to sort out primarily what to do via deliberation, which in principle is a process of sorting out intentions from desires being constrained by beliefs. Apart from deliberations there is yet another reasoning process in practical reason, the means-ends reasoning, which governs the 'how to' part of pursuing the intention, which is less emphasized in Bratman's original account outlined in (Bratman, 1987).

Very likely the popularity of the BDI framework is due to the fact that it can be elegantly translated into a computational model of rational behaviour of artificial agents. Selected formal implementations of the model will be presented in the following subsection. Agent-based approach to modelling human behaviour is underpinned by formal logic, which governs the AA's behaviour. In the below state-of-the-art review I will therefore pay special attention to logical frameworks applied to emulate human rationality in information systems.

5.4 Towards representing experiencing agents in information systems

In this section we will provide an overview of formalisms used for representing BDI agents in IS, then we will review already proposed frameworks for designing BDI agents that are *emotional*, with the purpose of identifying weaknesses of existing approaches, finally we will suggest direction towards including our account of experience, in particular the affective dimension of experience, into the BDI framework and provide arguments for why we believe this could constitute important enhancement of the framework from the point of view of efficiency of natural agent behaviour emulation and creating believable artificial agents.

5.4.1 Mainstream formal models of BDI agency

There have been many attempts to formalize rational agency among which BDI agent implementations are most common. Likewise, many BDI agency frameworks has been proposed so the state-of-the-art of this particular area is massive, full review of which falls beyond the scope of this thesis. However as we have chosen the BDI framework as a starting point for implementing experiential rational agency it is necessary to briefly introduce selected work in the area. We have rationed the scope of this introduction to two most popular frameworks: (i) BDI_CTL framework and (ii) KARO framework. Noteworthy, the first fully fledged BDI agency logic implementation was proposed by Cohen and Lavesque (Cohen and Levesque, 1990; Cohen and Levesque, 1991) which is based on an adapted linear temporal logic (LTL) enriched with primitive operators for belief and goal, operators providing for representation of actions (HAPPENSα and DONEα) and aprimitive action by an AA (ACT $i\alpha$: agent i is the actor of α), but this framework will not be presented here in more detail.

BDI_CTL framework One of the most common and cited logical implementation of the BDI model of agency is that proposed by Rao and Georgeff (Rao and Georgeff, 1991), which relies on the Bratman's theory of human action and a temporal logic CTL (Computation Tree Logic) (Emerson and Srinivasan, 1989) slightly adapted to the BDI requirements. The main novelty of this implementation compared to earlier endeavours to formalize the BDI theory is that for beliefs (B), desires (D), goals (G), which for Rao and Georgeff are consistent desires chosen by an agent, and intentions (I) are represented with a temporal structure of a time tree, with a single past and a branching time future, where an event is defined as moving along the time tree by a single point, referred to as a *situation*. B, G and I are therefore represented as sets of possible worlds, which result from the fact that agents do not have complete knowledge about the state of the world, *belief-*, *goal-* and *intention-accessible* worlds meaning respectively worlds believed by the agent to be possible, worlds representing AA's goals, worlds that an AA committed to realize. Importantly meanwhile intentions need to be consistent with goals, and goals consistent with beliefs, goals and intentions need not be closed under the beliefs of AA, due to which it is possible that AA commits to a goal that may result in side-effects that are not desired by AA, for which Rao and Georgeff provide the classical example of the visit at a dentist, which inevitably equals to the AA suffer pain if it wants to have his teeth fixed. Consequently an important feature of the model is that it follows tightly Bratman's insistence on the intentions being separate and hierarchically equal entities to desires and goals, not mere derivatives of the two. The implementation by Rao and Georgeff extends the CTL with two modal

operators: *optional* and *inevitable* and makes use of standard temporal operators for next, for eventually, for always and for until.

The framework proposed by Rao and Georgeff lets the Bratman's theory of human action be formalized so that it could be possible to formulate statements in a formal system about agent action however it does not provide an algorithm for complete agent's rationality, notably how an AA forms, maintains and revises the three sets, in particular how beliefs prompt goals selection and how an AA commits to them creating intentions. The authors underline that different types of agents may have different rules that govern these processes, which would determine their behavioural profiles. Some very basic examples are provided to demonstrate the usability of the proposed formalism.

KARO Framework The KARO approach has been proposed and developed by Van Linder, Van der Hoek and Meyer (Meyer and Hoek, 1995; Meyer, Hoek, and Linder, 1999; Hoek, Van Linder, and Meyer, 1999; Hoek and Wooldridge, 2003; Van Linder, Hoek, and Meyer, 1995; Van Linder, Meyer, and Van Der Hoek, 1997). KARO builds on propositional dynamic logic unlike Rao and Georgeff's BDI_CTL framework which relies on temporal logic. Van Linder et al. proposes a language of propositional dynamic logic augmented with modal operators for knowledge (K), belief (B), desire (D), and (A) denoting ability of an AA to perform an action. It is characteristic for this framework that it emphasizes on actions, action results and different mental states of an agent directly related to action, which may be a consequence or reason for tapping into propositional dynamic logic as the foundations for KARO agency logic. It moves further on defining agent's attitudes and situations resulting from agent acting upon intentions compared to theoretical BDI which emphasizes the deliberation and intentions more than means to ends reasoning. KARO achieves this by extending the core language with further additional operators which in total provide for a complete and extensive BDI agency logic. These additional operators include: *opportunity* ($\langle do_i\alpha\rangle\,\varphi$) denoting an opportunity for an agent to perform an action α with result φ; *practical possibility* (**PracPoss**) denoting possibility to do an action with respect to an assertion; *can* (**Can**) which stands for knowing to have the practical possibility to perform an action with respect to an assertion; *realizability*, i.e. the existence of a plan of actions that an agent has practical possibility to perform with respect to the assertion at hand; *goal* (**Goal**) which denotes an assertion that is desirable, not true yet, but realizable; *possibly intend* (**PossIntend**) to do an action with respect to an assertion (expressing that the agent can do the action with respect to the assertion of which he knows it to be a goal of his). Furthermore KARO includes special actions *commit* and *uncommit* to manipulate with the commitments of AAs. Naturally the framework covers the control of motivational attitudes such as wishes, goals and commitments of AA (Meyer, Hoek, and Linder, 1999; Segerberg, Meyer, and Kracht, 2009).

5.4.2 Review of existing emotional agency representations

The most promising and interesting directions of extension of the basic model of agent's behaviour in agent-based systems are: (1) incorporating in the AA other cognitive capabilities that go beyond logical reasoning, including feelings and emotions, (2) enabling social capabilities of AA. In this subsection I will focus on the former, in particular I will take a closer look at the emotional aspects of AA. The social capabilities of AA will not be tackled in this thesis as detail discussion of this theme is the ongoing work.

Given the discussion presented in earlier parts of the thesis on the role of emotions in human decision making as well as the way how experience has been defined and included in the adopted theory of human action, the importance of emotional dimension of an AA is self-evident. Still it is worth reviewing how computer scientists have been dealing with AA'a emotions so far and for which reason they find emotions an important subject of their studies. Several reasons could be given for why information systems, computer systems in particular, that rely on intelligent AAs could benefit from AAs having emotions. Sloman quite early (Sloman and Croucher, 1981), (Sloman, 1987) noted that emotions would play important role in AI systems, for they are immanent and integral part of human cognition, which is a fundamental benchmark for AI. As emotional states, he points out, arise from mechanisms required for coping intelligently in a complex enviroment there is no point in separating the emotion from cognition. Earlier Simon (Simon, 1967) suggested that emotions build up an "interruption mechanism" for general purpose information processor, as defined under his information processing theory of human cognition, which mechanism prompts the processor to interrupt operations as to respond to urgent needs in real-time.

It can be noticed that in principle there are two main reasons given for including emotions in the scope of the study of AA: (1) To enable believable emulation of human behaviour with AAs it is necessary to embrace emotional capabilities which are immanent to human cognition and action, believable behaviour of AAs is a key preoccupation of both AI scientists dealing with problem solving, socio-economic simulation and robotics as well as artists, animation and computer games specialists trying to engineer the illusion of real life, real living creatures (Bates, 1994); (2) as emotions play critical role in human cognition and decision making they must be central to human rationality at large, therefore including emotions in artificial reasoning systems should lead to efficiency gains, which is the main focus of software engineering. Meanwhile the former argument seems quite natural and intuitively acceptable, the latter seems a bit more controversial. It reads that we can make the artificial reasoning systems better by including emotions in the AA program.

This claim has been made indirectly by both Simon and Sloman, as well as explicitly by recent architects of emotional agent systems Meyer and Dastani:

> "We are interested in agents with emotions since we believe that emotions are important heuristics that can be used to define effective and efficient decision-making process." (Meyer, 2004; Dastani and Meyer, 2006),

and Jiang et. all:

> "We believe that (...) emotions can still be helpful for agents to make better decisions." (Jiang, Vidal, and Huhns, 2007)

Minsky is even more radical on this stance:

> "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions." (Minsky, 1988)

The need for emotional agents in systems that emulate human behaviour or imitate human behaviour in a more convincing and natural way, e.g. in expert systems used in studying social behaviour or in computer games respectively, is evident because there is sound psychological evidence that human behaviour, (linguistic) communication and social interaction is highly dependant on affect (see earlier sections and chapters). However I disagree that direct reference to emotion and affect is a promising direction in searching more optimal decision algorithms for artificial systems solving complex problems.

First off all, it must be remembered that affect is biological phenomena typical of biological beings, more precisely conscious biological beings. It seems convincing that emotion requires some form of self to be able to emerge and be sensed by this very self. Emotion entails subjective feelings of a leaving creature. It is confusing to talk about computer system's emotions and other affective states in first place, and we should insist on talking about affect representation or emulation in IS instead. Noteworthy, all above cited authors reserve from entering into philosophical debates on the possibility for an artificial system to be conscious or have emotions, rather they adopt a pragmatic, functionalistic view trying to see the practical benefits of looking at artificial systems as affect-capable entities. We think therefore that taking about agent's emotions in contexts other than representation of biological phenomena in information systems is confusing. We recognise that certain mechanisms for decision moderation that have "biological implementation" so to speak in the form of affective states, might be inspiring for designers of intelligent systems however it is possible but very unlikely that affect *per se* would contribute to the efficiency of problem solving systems. It has to be remembered that from evolutionary perspective affect is designed to help to take fast decision in the situation of reasoning capacity shortages and/or time constraints, not necessarily best decisions. It is true that affect ration human agent in the deliberative process and thus make it faster and effortless, however unlike artificial systems a human when involved in a deliberation start searching for

a satisfiable solution from already the best alternative that intuitively occurs to him, so one can afford limiting the search to a few first alternatives without major loss in choice efficiency.

Nonetheless if we want to construct systems that behave like humans, or create systems that emulate or predict natural agent individual or group behaviour, we need to put emotion into the loop. Below I will briefly describe the selected formal frameworks designed to represent and process emotions in agent-based information systems.

CTL based frameworks

EBDI - an architecture for emotional BDI agents. Jiang et. al (Jiang, Vidal, and Huhns, 2007) propose a generic architecture for agents with emotions which allows for separating the emotional mechanisms within an agent in parallel to practical reasoning. The main advantage of this architecture is that it allows to implement different specific emotional models, so regardless the selected theory of emotion that may be relevant for a given application this architecture could be applied. Another strong point is that it extends the BDI framework skilfully by including emotions, for instance (1) in EBDI emotions set the priority of desires and help decide intentions; (2) unlike original BDI the EBDI allows for intention-belief inconsistency which is handled by emotions, i.e. emotions are used as a tool to balance intention and belief as intention can influence belief indirectly through emotions; (3) complementary ways for acquiring beliefs are added: via contemplation and communication.

Similarly to other models EBDI considers emotions as a separate set. The state of the EBDI agent is defined as a tuple: $< E, B, D, I >$ so by the current sets of emotions, believes, desires and intentions. Furthermore the model defines in very broad terms how the emotions come into play with B, D and I, by specifying a few revision functions, specifically: (1) belief revision function: $brf : E \times I \times (B \cup B_p \cup B_m) \rightarrow B$, which shows that beliefs set is revised based on the emotions as well as perception (B_p denotes a set of belief candidates form preception) and communication messages (B_m denotes the set of possible belief candidates from communication); (2) two emotion update functions $euf1 : E \times I \times (B_p \cup B_m) \rightarrow E$ and $euf2 : E \times I \times B \rightarrow E$ for primary and secondary emotions respectively; as well as (3) a filter function $filter : E \times B \times D \times I \rightarrow I$ that revises the set of intentions to find the best option(s) for later execution (Jiang, Vidal, and Huhns, 2007). However the model does not specify exactly how emotions interfere with the other sets of BDI. It only proposes a generic process specifying at which stage of perception-deliberation-action emotions should be taken into account. The model provides no insights how emotions determine AA's actions leaving this part to the designers of particular implementation.

Emotional BDI - modelling agent's emotions with $\varepsilon_B DI$ logic The eBDI framework for modelling emotions in BDI agents proposed by Pereira, et. all (Pereira, Oliveira, and Moreira, 2006) relies on the work of Oliveira and Sarmento (Oliveira and Sarmento, 2002) who extended the OCC model of emotions by introducing a concept of *emotional valence* defined as "a subjective measure that relates the chances of an agent being able to fulfil its goals given a particular environment situation, its internal state and its set capabilities." This work was done with the intention to improve efficiency of AA's deliberation in dynamic worlds, so it falls within the research thread that recognizes emotions as important element of efficient reasoning. The definition of emotional valence stems from the observation that the OCC model primarily focuses on how cognitive processes leads to emergence of emotional states, i.e. the *eliciting conditions*, but it provides no insight in what is the role of emotions in enhancing the problem solving capacities of AA in complex environments.

Oliveira and Sarmento assert that fundamental role of emotion is the evaluation of the state of the world, so emotions provide for a mechanism that allows for rapid and automatic assessment of environmental conditions and its orchestration with internal states of AA. So the emotional valence is a sort of weights that reflect the AA's subjective estimation of chances of attaining a goal given the current state of environment and her mental state and capabilities. However the proposed model does not rule out whether this assessment is followed by an immediate response or is retained for later processing, neither it assumes if this response is a reactive or deliberative in nature. Furthermore in the model of emotional valence an AA evaluates the environment only in strict relation to its current goals for which reason it would not notice changes that are unrelated to his internal states, which is an apparent simplification, as it excludes situations in which AA's behaviour is influenced by external stimuli irrespective of a current deliberative process.

The way Oliveira and Sarmento understand and define emotional valance is very important as this concept is seemingly similar to the experiential model of agency as we proposed in Ryżko and Kaczmarek (Ryżko and Kaczmarek, 2011) and in earlier sections. However there is an important deference between these two approaches, which is the way each approach understands emotions ontologically. Oliveira and Saramento reduce emotional mechanisms to conscious rational assessment of the relationship between the environment and internal mental states, primarily goals. This is an evident influence of appraisal theory of emotion which we have refuted as incomplete. Oliveira and Sarmento see emotions as providing "an automatic and quick way of evaluating the environment and the internal state of the agent in respect to its own goals". This evaluation is directed both outwards to the environment and inwards towards the internal states such as beliefs, goals, action sets, etc.. Emotional valence answers the questions how are, negative or positive and to what degree, both environmental conditions and current plans, knowledge to the specific goals of an AA. Meanwhile our account first distinguishes between emotions as programmed responses

to specific stimuli and other affective, feeling states which colour intentional contents of perceptions. These are remembered and reconstructed from memory together with relevant intentional contents, i.e. instances of knowledge.

As explained in chapter 3 neurocognitive research suggests that at the foundational level emotional affective states are biologically and neurologically independent of conscious, rational information processing in the cortical parts of the brain. Furthermore the study of some emotional neurological cerebral circuits, such as fear system, show that an emotional reaction to an external stimulus precedes the conscious appreciation and modulation of the perception (LeDoux, 2000). This makes it unlikely that the emotional valence as defined above well represents the way emotions and its relation to the conscious deliberation. Instead, as we have proposed, emotional valence and intensity colour agent's knowledge forming part of unified field of conscious experience. Assigning emotional valence and intensity to atomic intentional states, represented by propositions also allows for better modelling of emotional influence on deliberation across time, as affective states are assigned to and remembered by the agent with a corresponding intentional state. Under Oliveira's and Saramento's model the emotional states are recorded in short- and long-term memory in the form of *Valence Vectors* $< V, I, E, G >$ where I stands for input from internal sources, E - input from external sources, V - a valence measure regarding a given goal G, which is useful from an engineering point of view, but would require that an agent be highly consistent and disciplined in storing and recalling its emotional states. However the most significant limitation is that such definition of emotional states in AA does not allow for handling such phenomena as orchestrating conflicting emotions and associative proliferation of emotional states, for instance if an agent for some reason experiences strong emotions in relation to an object of some particular property, e.g. red colour, agent's behaviour with respect to other red objects is likely to be affected even if this new object has not much to do with the object being the source of emotional arousal.

Inspired by this model of emotion Pereira et.al have proposed an elaborate formalism based on BDICTL logic and earlier work of Rao and Georgeff, with elements of KARA framework for representing emotional states in BDI agents. They extended the DBICTL logic with elements of KARO, namely operator for representing *capability*: Cap, Can, Cannot and EffCap (effective capability); action representation from Propositional Dynamic Logic (PDL), modal operator Fund for *fundamental desire*, modal operators Res, Needs, Available, Saved and atomic actions get(r), save(r) and free(r) for resources management; and operators AtRisk, PossATRisk, and Safe to handle risky *situations*, finally standard temporal operators of CTL allow for considering emotions changing overtime, however authors do not make much use of it as far as emotional dynamics are concerned but do take advantage of it in considering risk. With this apparatus at hand three exemplary emotions: fear, anxiety and self-confidence are defined and some examples how these could influence

AA's behaviour are given. The way how the framework deals with threats as well as combining action execution and time with resource management by AA allows for formulation of alternative eliciting conditions for emotions such as fear, anxiety and allows for embracing more emotional states such as self-confidence, which is a clear enhancement compared to the framework proposed by Meyer. However the model does not leads to a major breakthrough in emotion handling in BDI agent systems, rather allows for incremental improvements or alternative formulations that may be more adequate for some particular applications. The model better relates to the influence of emotions on agents behaviour as the tableau construction of BDICTL has been expanded by formulas referring to fundamental desires (Fund) still the examples of application did not show convincingly substantial potential of this framework in enhancing overall AA's performance.

EL logic of emotion based on OCC model of emotion Adam in her thesis (Adam, 2007) undertakes to come up with a generic formal model of emotions as a reply to observed heterogeneity of approaches in modelling emotions by AI researchers, concluded from her review of relevant literature. Indeed different AI researchers have undertaken separate endeavours aimed at proposing a plausible model of emotion representation in agent-based systems (cf. previous sections). Often the frameworks overlap and some kind of embracing alignment or unification of this theories would be welcome. Adam decides to make this unification on the basis on one, subjectively selected theory of emotion and agent rationality model, namely the OCC theory of emotions and classical BDI logical implementations.

The entire framework for emotion representation proposed by Adam is very elegant, also due to the fact that it relies on the proven logical framework for representing BDI agents, as she builds on the previous research of Rao and Georgeff, and Wooldridge. She provides for 20 formal definitions for 20 emotional states as proposed by the OCC model by means of standard operators for believe and desire Bel, Des and six additional modal operators provided by the applied formal system, i.e. $Expect$, $Prob$, $Done$, Idl, $Happens$, $After$.

However this representation inherits the weaknesses of OCC model, which though widely accepted by AI community for its simplicity is not a most respected theory of emotion in psychology, for which reason examples and intuitions about emotions provided by Adams and deduced form proposed axioms are not very convincing. For instance, Theorem 12 claims the principle of non simultaneity of hope and fear, i.e. $\vdash \neg(Hope_i\varphi \wedge Fear_i)$ which is a consequence of the definitions of Hope and Fear and the principle of consistency of expectations, because *hope* implies expectation of φ while *fear* expectation of $\neg\varphi$ (Adam, 2007, p. 129). However we discharge the claim that hope is never accompanied by fear, on the contrary we claim hope is derivative of fear. Hope can be seen as expectation to avoid some negative state of affairs that we fear about. So once we hope at the same time we fear about our

expectation not to be met. As we fear we want to avoid the state of affairs that cause fear, therefore we want to release the fear by hoping, but this does not stop us from forgetting our fear, or fear is there all the time but it is *overshadowed*, so to speak, not *replaced* by the hope. This also relates to the known odds between OCC model and Lazarus' theory of *hope* emotion (Lazarus, 1991) according to which hope is the emotional state that comes about when something negative is expected to happen but nonetheless one believes it may turn out less negative in the end; on the contrary the OCC model accounts for hope as arising when something positive is expected but it is merely likely not sure.

The weakness of the Adam's model in this case is that it does not allow for coexistence of all types of emotions, taking the approach that contradictory emotions replace each other not that they coexist but one simply dominates, or outweighs the others, which is much more intuitively close to such a fluent phenomena as emotions.

Another inherited flaw is the definition of *joy* caused by φ as conjunction of a believe that φ and desire that φ, which in notation adopted by Adam is: $Joy_i\varphi = Bel_i\varphi \wedge Des_i\varphi$. In other words an AA_i is happy when it realizes she has satisfied her desire, likewise AA is in distress when she realizes that something undesired happened. This approach seems already slightly broader than that of Meyer's (cf. previous section) which made joy dependent on successful achievement of goals by an AA, nevertheless still too narrow to embrace a situation when an AA becomes happy as a result of a sole revision of beliefs, a good example of which is a joke or funny situation that an AA senses. To give an example if an agent is presented with an amusing cartoon he will evidently become happy meanwhile he has not achieved any goal nor satisfied any desire.

Likewise, as in previously discussed frameworks under Adam's account proliferation of affective states across mental states is impossible, there is no place for affective association between believes, no affective serendipity which could importantly alter agent immediate or future behaviour. For instance, let us consider a case of an AA emulating consumer behaviour exposed to a commercial communication, e.g. an advertisement on the radio. The advertisement modifies AA's B (beliefs) set, but likely the commercial would carry heavy emotional load shaping the affective intensity and valence of the AA's experience. Now agent's mind would capture the experiential state encapsulating both new knowledge from the event (the facts, propositional contents conveyed by the commercial) as well as the corresponding affective load. Through spontaneous association in the process of deliberation the agent should be able to associate the affective and intentional contents of the advertisement once confronted with the advertised product as well as in similar affective context. Such affective interplay within (B) believes set which in turn can bias consumer behaviour, i.e. prompt the revision of B, G or I sets is not supported by any of the reviewed accounts.

Yet again though the framework appears to be sound there is a clear lack of strong examples that would prove the framework unconstrained applicability to real problem solving. The proposed axiomatization of emotional dimension of AA leads to theorems that claim emotion properties that are counter-intuitive or do not appear likely if we take a step back and look from a wider, practical perspective or from the standpoint of alternative theory of human affect. It introduces consistency rules that oversimplify emotional dimension of AA rationality, which is acceptable as a first step, however there are no clear perspectives provided on how to make this model more inclusive. Important limitation of the model is that it does not allow for coexistence of conflicting emotional states and it presents overly simplified and somewhat naive exemplification of AA' emotional life.

KARO based frameworks

LEA - Logic of Emotional Agents. In a series of papers Mayer and his colleagues (Meyer, 2004), (Dastani and Meyer, 2006), (Steunebrink, Dastani, and Meyer, 2007) have introduced a logical system for representing emotional agents based on the KARO logic which they called LEA (Logic of Emotial Agents). The development of this framework was started by Meyer (Meyer, 2004) who relying on the psychological evidence recognised that emotions significantly influence behaviour via creating attitudes that are responsible for handling AA's goals and intentions and proposed the extension to the KARO logic developed earlier together with van Linder, to cater for emotional attitudes. As emotional states were defined as attitudes towards goal maintenance and execution they were represented under LEA as fluents, predicates that can change over time. Each fluent had two axioms, one specifying under which condition the emotion arises, the second how the emotion affects AA's behaviour. Consequently these axioms are constraints on the deliberation and execution of startegies that an AA applies. Furthermore Meyer considered 4 basic emotions: happiness, sadness, anger and fear and proposed a couple of axioms governing these emotions in LEA, which represented some very basic heuristics such as: (1) AA becomes happy when she succeeds in achieving subgoals, and once happy an AA is more persistent in executing its intention; (2) On the contrary, an AA that fails to attain her subgoals gets sad which prompts her to abandon the chosen course of action and revison of plans, intentions or goals; (3) if an active plan of an AA collapses she gets frustrated and angry which makes her revise her believed capabilities or defer the execution until she is capable of achieving the set goal or persists in the execution of her intentions with alternative means, which would depend on this AA's behavioural profile; (4) Finally an AA can become fearful when she confronts conflicting goals, after which she is likely to revise her plans looking more carefully on how environment changes to avoid conflicts in the future.

The main downsize of the approach taken by Meyer and his colleagues is that it maintains the emotional states as ontologically separate facts, irrespectively of any other mental states of AA. We consider this approach mistaken. According to very plausible theories of consciousness, conscious experience is gained within a unified field, so any emotion is tightly bound with other mental contents that accompany affect, or better put, that build up the given affective state. Therefore we argue that instead of maintaining emotions as separate beliefs it would be better to refer to experience at large represented as affected intentional states. In case of formalisms proposed by Meyer et. al we could consider all beliefs to be affected, i.e. to have a permanently assigned affective value that would become a property of a given belief, remembered by an AA together with belief propositional content.

A minor comment should be made on the fact that LEA proposes only exemplary and very simplified axiomatization of AA's emotions, leaving it open for particular implementation, which in principle is a correct approach however by now there have been no LEA implementations that could prove the framework to have practical value.

Miscellaneous approaches

Emotional gauges Padgham and Taylor (Padgham and Taylor, 1997) introduce a simplified model of emotions in which emotions are grouped in pairs of opposites (e.g. love v.s. hage, pride vs. shame) and are represented by a gauge with a neutral point about which the emotions fluctuate in positive or negative direction. The fluctuations of emotional gauges are caused by events and passing time. Give the time decay rate of emotional state is represented by a function $D_A(e)$ Therefore for each particular agent A the following function can be defined for determining the values of the emotional guauges:

$$V_{A_{t+d}}(e) = V_{A_t}(e) + F([events]_t^{t+d}, D_A(e,d)) \tag{5.1}$$

The moment the emotional gauge passes a threashold defined for agent A emotions start to take effect on agent's cognition and behaviour. This allowed the authors to define agent personality as a tuple: $Phi_A =< M_A, N_A, P_A, D_A >$, where M_A is the set of motivational concerns for agent A, $P_A(e)$ and $N_A(e)$ are the functions representing negative and positive assertion thresholds for the gauges, D_A the decay function and the $D_A(e)$ the set of motivational concerns for agent A, which govern the agent's desires.

This simple model is useful for representing emotions in an agent system in simple applications such as modelling characters in computer games. However it has some evident limitations. First of all it relies on a simplified theory of emotions, which may be questioned from the emotion theoretical point of view, for instance insofar as classifying the emotions and pairing them on the principle of opposites.

Secondly it does not explain the way how emotions motivate actions. Nevertheless it elegantly models the personality as a behavioural pattern that reflects how an AA manages emotions. It also allows to represent the intensity (degree) of emotions. The simplicity of the model is appealing and the fact that it is strongly influenced by the dimensional emotion theory makes it an interesting alternative to frameworks blindly following outdated appraisal accounts.

Modelling Emotions with Multidimensional Logic. MEML is a model proposed by Gershenson (Gershenson, 1999) that takes three sets of emotions represented by a bidimensional logic variable each: love/hate, joy/grief, and happy/sadness by leanear combinations of which more complex emotions such as pride, conformity, boredom and others. Gershenson defines 20 exemplary emotional states all together. The model does not only provide for the fuzziness of the intensity and valence of the given emotion which can take values from within [-1;1] but also defines a fuzzy frontier the basic three emotional variables, so that any emotional state can be defined as a vector in 3-dimensional space where on x and y axes the valence of negative and positive dimension is represented and on the z axis the three emotional variables (love/hate, joy/grief, happiness/sadness) are delineated, therefore we can define an emotional state that is on any plain between for instance happiness/sadness and love/hate.

Although the idea to represent emotions with multidimensional logic is appealing, as it allows to reflect the fuzziness of emotional phenomena as embraced by dimensionalist theories of emotion (see earlier sections), the visible problem is how to make such representation fit any plausible theory of affective agency. In other word intuitively we know that emotions are ill-structured, however until we manage to structure them it is difficult to make any use of them, map them onto action. The taxonomy of emotions represented with the formalism proposed by Gershenson confirms the issue, as the author himself suggest that it is merely an exemplary representation which indeed seems arbitrary, not supported by nor linked to any psychological theory of affect bound behaviour.

5.4.3 Experiencing BDI agents

Before we propose ways how our approach to experience representation could enhance the BDI agency framework let us summarize the main limitations of the existing frameworks for capturing DCI agent's affective states:

1. As a general rule emotional agency frameworks are based on outdated or flawed emotion theories for which reasons despite elegance of formalization they foster application of limited usability;

2. Both formal constraints and dependence on imperfect theoretical accounts of affect renders the frameworks limited in expressiveness and incapable of comprehensive mapping of affective phenomena onto agent behaviour. Formal constraints dictated by requirement of completeness and soundness of the formal systems renders the limited flexibility in treating the fuzziness of affective phenomena meanwhile lack of sound theories of affect driven action makes for the frameworks impossible to provide results of practical value. The practical examples used to prove the application relevance of the frameworks are not convincing and present folk's knowledge about relation of emotion to behaviour;

3. All frameworks in principle confuse emotions as hard-wired programmes and emotional feelings, they do not incorporate contemporary knowledge about affect, affect systems and affective feelings. In principles there is little interdisciplinary thinking behind the frameworks, which are made legitimate by mere reference to old psychological accounts with little or no critical reflection on their validity by the engineers of the formal frameworks.

4. Meanwhile impact of emotional states, understood as programmes, is well captured by mainstream KARO or CTL frameworks the impact of affect on believes set is underestimated. Emotional states are treated as states the agent is consciously aware of, as emotions are deduced from consciously appreciated mental states: desires, intentions and believes.

5. None of the frameworks known to us is capable of addressing unconscious cognitive or affective processes in deliberation or means to ends reasoning, nor any other direct influence on agent's behaviour

6. Temporal dimension is applied only insofar as the sequence of intentions maturity and action implementation stages are concerned, not affecting the believes set. None of the framework addresses the fact that believes and associated mental states, including affect decay with time

7. None of the frameworks properly addresses the memory which may be a weakness inherited from the BDI framework

8. The available frameworks tend to mix old concepts of emotions with contemporary, for instance they talk about love, hate, fear sadness, pride, meanwhile some theories talk only about fear, we are not sure how many emotions there are and how to classify them.

9. There is evident lack of "unification framework" that would embrace affective phenomena inclusively and allow for decent mapping of affective states onto behavioural consequences. This is partly due to the fact that we are still at the

beginning of understanding how emotions influence behaviour, there is no one powerful theory such as the BDI or intentionality theory of human mind.

10. Importantly, as we emphasized earlier in chapter 3 there are limitations in linguistic capacities to express affective states, meanwhile all the frameworks rely on prepositional language in representing agent's mental states regardless their nature B, I, D or G, none is capable of describing and representing affective phenomena in neuronal terms, which becomes a common standard for considering affect in contemporary neuroscience and science at large.

11.

Given the above the usefulness of the above frameworks is very limited. They may be applied with considerable success in creating believable virtual characters for video games and other entertainment and arts multimedia content, with moderate success in software engineering providing some improvements to agent-based systems, and are too weak to provide basis for satisfactory emulation and prediction of behaviour of natural agents in information systems, likewise for representing experiential phenomena in expert systems for support in decision making, for instance in customer relationship/experience management, likely rendering worst results for individual behaviour then emulating group behaviour dynamics.

Now, let us consider how the experience representation framework could enhance the BDI emotional agency frameworks.

First of all it must be noted that our approach to experience representation could be applied primarily to enhanced representation of the B (beliefs) set. We shall recognize that the BDI theory of practical reason as proposed by Bratman is one of the most advanced AI friendly accounts of agency. Bratman's contribution to understanding how human take action is highly valuable. However it pretty much neglects the role of affect in practical reason. It may be because this account is firmly rooted in cognitivist tradition. This constitutes a constraint for any BDI agency formalization right from the start. Our account of experience understood as affected knowledge could patch this evident gap. This would be achieved in two ways: (1) by providing a framework for affect control in the framework, in particular as far as modulating memory volatility processes as far as both knowledge and affective dimension of experience is concerned, (2) for orchestrating affect influence on behaviour.

We could risk formulating a simplification that there are two basic ways in which affect influences agent's action: (1) via immediate impact on behaviour when one of the known emotional systems is activated, i.e. when emotion program sets off, altering immediately behaviour of an agent, interfering its normal deliberation and means-ends reasoning processes, (2) indirectly influencing deliberation process via impacting believes set. These two main types of influence overlap in situations when

recalled affective states associated with processed beliefs invoke affective arousal intensive enough to trigger fully-fledged emotion. This is we believe one of the biggest advantages that our account could bring to BDI framework. In wider terms this is about providing for a proper temporal dimension of experience and modelling experience as learning associative process, including memory processes.

Furthermore, as the framework does not *a priori* impose any particular emotion representation system it could flexibly embrace contemporary models of affect that consider affective states in neuronal or some other terms, for instance it could incorporate a crafted affect coding system for different identified by contemporary neuroscience affective states represented for instance with multidimensional vectors, where vector space could be delineated by different brain systems that are activated or inhibited under given affective state.

Moreover, as detailed control of affective state of an AA would be provided the BDI framework could be enriched with more sophisticated rules definition for mapping these states onto behavioural consequences. Affect could influence deliberation in variety of ways, by highlighting dynamically importance of certain desires and believes on the expense of another. Intention formulation could also be modulated by the affective state of the agent at any point in time. Reconstruction of past mental states and their amalgamation with current new perceptions, as unified field of consciousness account dictates, could be emulated. With enhanced affective dimension DBI framework could better address motivational aspects of intention formulation. Likely, the gap of the free will could be bridged, even if only with provisional approximations, based on the affective value estimation of alternative states.

In order to validate above theoretical assumptions empirical experimentation are necessary. It is a common problem of all BDI formalization as Meyer rightly concludes in his recent review of formal logic implementation of BDI agency:

> "To this day there is still a gap between theory and practice" (Segerberg, Meyer, and Kracht, 2009)

Also in our case empirical verification of our experience theory constitutes ongoing work.

5.5 Conclusion

Simon rightly pointed out that it should be the goal of AI to design agents that overcome human limitations while displaying all their strengths (Simon, 1967). Sloman in turn speculated that certain limitations of human agents, resulting in failing to achieve fitness optima on some occasions, may be inevitable consequences of their strengths resulting in human intelligence (Sloman, 1999).

Although we see little potential of experience representation frameworks, such as proposed herewith or other, in fostering the creation of artificial intelligence or delivering improved machine learning techniques or software engineering paradigms, there is a vivid need for improved frameworks that could capture subjective affective phenomena in information systems for applications in human-machine interfaces, emulation of individual and group behaviour emulation for application in social sciences (incl. multi-agent simulation frameworks), emulation of consumer behaviour and representation of customer experience in Customer Experience Management systems and marketing decision support systems, as well as creation of believable virtual characters in games and entertainment industry. We hope that the account of experience we have proposed could be successfully applied in the aforementioned areas contributing to increased efficiency and usability of information systems of various classes.

Chapter 6

Conclusion

6.1 Conclusion of findings

The objective of this thesis was to theoretically verify the proposition that experience can be represented in information systems and that such representation is vital to enable emulation of behaviour of natural agents in these kind of artificial, symbolic systems.

To achieve this the nature of conscious experience and its role in the behaviour of mindful organisms, foremost human beings, has been studied in-depth from a wide range of disciplinary perspectives including information theory, philosophy of mind, neuroscience, economics and psychology. The literature study in these areas has allowed us to identify the constituting elements of conscious experience and come up with the definition of experience amenable to formalisation and consequently propose a general-purpose framework for experience representation in information systems.

The reference point for the creation of the framework were the findings of philosophers, neuroscientists and economists such as Searle, Davidson, Bratman, Nagel, Tye, Damasio, Panksepp, Ledoux, Knutson, Loewenstein. Their essence was the appreciation of the role of subjective dimension of experience related to feelings, affect and emotion in shaping human mind, self, choice and behaviour at large accompanied by objection to both dualism and reductionism.

The evidence from both contemporary philosophy of mind and neuroscience has appeared to us sufficient, to assert that affect plays a central role in human and animal decision making. Also the reported results from an empirical study of our own strenghten this assertion and allowed us to better understand and experience the nature of the effect of affect on rational judgement. Furthermore the findings and theories by the above mentioned scientists, supported by convincing evidence, suggest that affect, the subjective feelings it invokes, is fundamental to constituting conscious mind and self. This has let us realize that there is an important aspect of human and animal experience that knowledge does not embrace. The missing subjective component of experience is affect that forms and integral part of natural

agent's experience, and determines, or at least impacts profoundly the behaviour of natural agents. Meanwhile, information science has focused on knowledge as the ultimate concept for representing mental phenomena, limiting thus expressiveness of the proposed frameworks with regard to satisfactory emulation of reasoning and consequently behaviour of natural agents.

We have proposed that relating affect to knowledge results in a satisfactory approximation of a broader concept: *experience* that more inclusively embraces mental phenomena with regard to both their contents and quality. Moreover we have demonstrated how the subjective dimension of experience, classified as affect, could be disentangled and represented building on neuroscientific account of affect and emotions which regards them not only as private, subjective epiphenomenal entities but rather qualities of conscious mental states that have neurological correlates in the brain which can be objectively studied. By identifying affective correlates of intentional contents of states of mind, which build up knowledge, we can exploit the broader concept of experience for the purpose of more accurate emulation of natural agents' reasoning and behaviour in information systems.

Based on this we have postulated that any intentional state, which is a representation of external world in the mind, has an affective value, which is characterised by valence (positive or negative), intensity (arousal level) and mode (affective state kind), which has implications on agent's behavioural response and is an integrated component of agent's rationality.

Consequently, it has been possible to propose a general framework for representing thus defined experience and relate it to the mainstream approaches to modelling rationality and emotions of rational agents in information systems, which we recapitulate below.

The framework consists in a general purpose definition of experience understood as remembered intentional states of mind. Formally, experience is defined as a pair of sets K and A, where K represents knowledge, that contents of remembered intentional states of mind, or intentional contents of experience, whereas A represents affect, i.e. the subjective qualitative component of experience, therefore:

$$E_J =< K, A >, \text{ where, } E \text{ - experience of agent J.}$$

Further we have defined a function mapping intentional content into affective state

$$f : K \to A$$

The element representing affective component of experience – A has been further defined as a set of tripples:

$A = \{< v, i, m >: v \in V, i \in I, m \in M\}$, where V – valence, I – intensity, M – mode, and M is a k-combination of C where C – n-element set of core affects and $k < n$ – the number of core affects involved in a compound affect.

In line with the contemporary neurocognitive theories of emotion and affect (Panksepp, 2005), affective component of an experiential state is characterized by *valence*, as a mindful organism can always discriminate between wanted, unwanted or neutral subjective states, *intensity* as there can be degrees to which these states are wanted or not, and finally they are characterised by a *mode* as there are neurologically recognised emotions each corresponding to the activation of a particular neural circuit in the brain, or a few circuits at a time. In line with contemporary neuroscientific theories of affect, we have distinguished between low-level, primordial affective states, i.e. *core affects*, like fear, lust, etc, and compound or high-level affects that can involve a combination of core affects, which is why we have defined M - mode, as k-combination of the set C, where k is any integer such that $k \in < 1; n >$.

While working to confirm the main thesis proposition and the development of the above outlined experience representation framework we have managed to achieve the following additional results:

1. We have made a thorough review of the definitions and understanding of experience across a wide variety of fields including psychology, philosophy, neuroscience, information science, affective computing and economics, which allowed for explanation of experience which is considered an ill-structured term.

2. We have came up with our own, formal definition of experience, which allows for representing experience in information systems.

3. We have came up with our own selection of interdisciplinary evidence for that affect, which stands behind the subjective component of experience is intrinsically related to rational behaviour, pointing to the necessity of redefinition of the traditional notion of rationality, which casts new light on the concept of experience.

4. We have carried out an empirical study that confirmed that affective stimuli can alter rational judgements.

5. We have demonstrated how the developed experience representation framework could enrich the mainstream model for representing natural agent rationality in information systems the so called BDI rationality model based on Bratman's account of practical reason.

6. We have also showed how the framework could be applied to modelling experience in an information system responsible for collecting, storing, retrieval and processing of information on experiential states of consumers under the so called Customer Experience Management systems, which purpose is to provide information and decision support to marketing and management decision makers responsible for customer relationship and customer satisfaction management, mostly in business organizations.

The above briefed results can be utilised both in our further investigations as well as can be utilised by other researchers in the area of information systems, in particular information systems which objective is to collect, store and process information about experience of natural agents. They can be embraced by a wide range of areas of application, for instance: user-centred information retrieval systems, decision support systems (e.g. customer experience management systems), agent-based simulation systems emulating behaviour of natural agents, human-computer interactions, recommender systems, user profiling, crisis management and decision support systems.

6.2 Limitations of experience representation methods

This thesis makes some progress in broadening the capacities of informations systems for a more complete representation of human knowledge and experience. Still, as it relies on the knowledge definitions as formulated by KR field it inherits important limitations. These limitations are to large extent inherited from the limitation of language in representing the human experience. Though this problem has been discussed in earlier chapters let us recapitulate here the most important of them:

1. Expressiveness of language is limited, this is why progress of knowledge in the scientific sense is done mostly by defining new terms, as clearly the scientific discovery is not about creating new reality but aptly describing its existing states with language. Still this proves that naming natural phenomena is hard and takes significant time and effort.

2. More importantly, language is evidently not the native carrier of knowledge for the human brain/mind to which many arguments can be provided: (i) children development (ii) animal intentionalistic capacities (iii) fairly recent cross-language comparative studies reviewed by Malt and Wolff (Malt and Wolff, 2010) reveal that different language communities tend to map human experiencess onto words differently which cannot be explain by mere differences in behavioural practices and culture. Surprisingly it has turned out that the diversity in ways how different language communities map concepts to language is much richer than expected. Malt and Wolff assume that there are only few or no domains of human experience in which the vocabulary of the domain maps cleanly onto one another across languages even for culturally closely related language communities. This shows that human cognitive architecture is not necessarily a straightforward and universal concept-word mapping apparatus. This may suggest that people in general differ in ways of mapping their experience onto language.

3. Neurocognitive studies in memory and imagination show that the native format of manipulating concepts in human brain is not language (Damasio, 2010; Kosslyn, 1996). The CDZ model proposed by Damasio (Damasio, 2010) and Kosslyn arguments on imagination reveal that the brain stores and processes memories in the formats that are pictorial in character (visual, auditory, tactual, etc.) rather such that remind linguistic descriptions.

The above important limitations of language, in particular with regard to describing affective phenomena stand for the intrinsic weakness of any language-based experience representation framework. Affect is not a language based phenomena, it has more primordial, non-linguistic forms of representations in a natural agent so any representation of the experiential phenomena in language-based information systems can only be an approximation. However we have argued that such approximations can be satisfactory for a number of real-life applications and can have important impact on information science, economics and management.

6.3 Ongoing and future research work

This thesis represents only the first attempt to approach the problem of experiential or affective theory of information, experience representation, processing and retrieval in non-biological information systems, and experiencing rational agents.

There are ample opportunities of advancement of this work, the following are some most promising and challenging directions of further research to be undertaken by the author:

1. Continuation of the development of the experience representation framework so that it better addressed temporal, volatile, subjective and dynamic character of natural agents' experience.

2. Construction of behavioural models that would map experience onto agent rationality, i.e. further elaboration of the *rational experiencing agent* model and behaviour emulation frameworks.

3. Coming up with efficient, user-centred and usable experience assessment and measurement methods, including both interfacing tools for self-assessment as well as inference methods for experience assessment from body parameters, including brain activity measurements.

4. Experimenting with real-life data for further validation and improvement of theoretical models, frameworks and assumptions.

5. Analysis of the social dimension of experience and the inclusion of social layer into the representation framework for modelling individual experience, conse-

quently modelling social experience and its relation to social phenomena such as morality and culture and eventually mapping these onto group behaviour.

6. Constructing social experiencing agents interacting in highly dynamic social context.

7. Investigating further and proposing improved affect representation frameworks, adapting these to new evidence from affective neuroscience and cognitive neuroscience at large, which is currently one of the most dynamically developing area of science.

8. Investigating non-linguistic forms of intentionality of mindbody, both in the inward direction (subjective intentionalistic states) and outward direction (objective intentionalistic states) with the purpose to come up with better forms and methods for expressing experience outside of a living organism.

Appendix

Contents of the survey referred to in section 3.5[1]

Survey on the quality of Polish wiki services

Thank you for participating in my survey! This will help me complete my research project, for which I am very grateful!
Please find below a short instruction:

1. The purpose of this survey is to ASSESS THE QUALITY OF ARTICLES in Polish wiki-type online services with an INTERSUBJECTIVE approach.
2. You have been assigned with one RANDOMLY CHOSEN entry from the online wiki-based cookbook WikiKuchnia.org.
3. Below you will find a few introductory questions, necessary to CALIBRATE THE RESULTS.
4. On the next page of this survey you will find a link to the selected wiki entry. Click that link (it should open in a new window). Read the article and then return to this survey.
5. This should not take more than 10 minutes.
6. This survey is ANONYMOUS.

Thank you very much for taking part in this research experiment.

[1]The survey questionnaire was originally authored in Polish. The provided translation has been prepared solely for the purpose of this appendix.

Introductory questions

Your gender female / male

Your age 18–24 / 25–43 / 35–44 / 45–54 / 55–64 / 65 and more

I like spending free evenings. . .

	(almost) always	often	rarely	(almost) never
at home at my PC/TV/reading a book	◇	◇	◇	◇
in a club/cafe	◇	◇	◇	◇
at a restaurant or hosted dinner	◇	◇	◇	◇
in a cinema/theatre	◇	◇	◇	◇
on physical exercise	◇	◇	◇	◇

Select which of these sentences correctly describes your attitude towards food and cuisine

☐ I pay a lot of attention to what I eat. I select dishes and recipies carefully.

☐ I like cooking.

☐ When going to a restaurant I care about what cuisine they serve and check out opinions beforehand.

☐ I like to experiment with food. I enjoy exotic meals and nonstandard flavours.

☐ I eat to get calories and do not care much about what I eat.

☐ None of the above.

What is your attitude towards the following ingredients used in meals?

	very much like	like	neutral	dislike	strongly dislike
minced pork	◇	◇	◇	◇	◇
minced beef	◇	◇	◇	◇	◇
onion	◇	◇	◇	◇	◇
oil	◇	◇	◇	◇	◇
flour	◇	◇	◇	◇	◇
yeast	◇	◇	◇	◇	◇
full-fat milk	◇	◇	◇	◇	◇
margarine	◇	◇	◇	◇	◇
egg yolk	◇	◇	◇	◇	◇
whipped egg white	◇	◇	◇	◇	◇
sugar	◇	◇	◇	◇	◇
salt	◇	◇	◇	◇	◇
pepper	◇	◇	◇	◇	◇

Are you hungry right now?

yes, very / yes, relatively / no, not really / no, I'm quite full.

Now read the article!

Open the link provided here to an WikiKuchnia.org article with a recipe. Read the article carefully and return to the survey. **[link]**

Main part of the Survey

Please fill this questionaire (THIS IS THE LAST STEP!) and do keep in mind that:
1. we want to hear your subjective opinion,
2. it is your first impression that counts,
2. IT IS FORBIDDEN to look up further info about the article trying to verify it.

Did you recognise that meal? yes / no

Have you eaten it before? yes / no

Do you feel like eating/preparing it after having read the article?
yes / no

Would you recommend this recipe to a friend? yes / no

Imagine that you are at a restaurant and find this dish in the menu among other dishes at the same price, dishes you know well and like. Would you order this dish? yes / no

Even if you have never tasted it before, try to imagine its taste. . . How would you rate the taste?
very much like / like / neutral / dislike / strongly dislike.

How would you rate the quality of the article you have just read?

	very good	good	poor	very poor
clearness	◇	◇	◇	◇
conciseness	◇	◇	◇	◇
structure	◇	◇	◇	◇
language	◇	◇	◇	◇
conformity (only if you knew the recipe)	◇	◇	◇	◇
visual aspects	◇	◇	◇	◇

Would you recommend WikiKuchnia.org to a friend? yes / no

Bibliography

Aaronson, S. (2007). "The Limits of Quantum Computers". In: *Computer Science – Theory and Applications*. Ed. by V. Diekert, M. Volkov, and A. Voronkov. Vol. 4649. Lecture Notes in Computer Science, p. 4.

Adam, C. (July 2007). "Emotions: from psychological theories to logical formalization and implementation in a BDI agent". PhD thesis. Université Paul Sabatier. Institut de Recherche en Informatique de Toulouse.

Allais, M. (1953). "Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine". In: *Econometrica* 21.4, pp. 503–546.

Allen, R.G.D. and J.R. Hicks (1934). "A Reconsideration of the Theory of Value". In: *Economica* 1.1, pp. 52–76.

Anscombe, G.E.M. (1957). *Intention*. Harvard University Press.

Aristotle and J. Sachs (2001). *Aristotle's On the soul: and, On memory and recollection*. Green Lion Press.

Armstrong, D.M. (1979). "Three types of consciousness". In: *Ciba Foundation Symposium*. 69, p. 235.

Barrett, L.F. (2006). "Are emotions natural kinds?" In: *Perspectives on Psychological Science* 1.1, p. 28.

Bartee, E.M. (1973). "A holistic view of problem solving". In: *Management Science* 20.4, pp. 439–448.

Bates, J. (1994). "The role of emotion in believable agents". In: *Communications of the ACM*.

Bauman, Z. (2011). "Duch i ciało na rynku - duchowość na sprzedaż". In: *Miesięcznik Znak* 7-8.674-675, pp. 17–22.

Begg, D., S. Fischer, and R. Dornbusch (1993). *Ekonomia*.

Belkin, N.J. (1996). "Intelligent information retrieval: Whose intelligence". In: *ISI* 96, pp. 25–31.

Benartzi, S. and R.H. Thaler (1995). "Myopic loss aversion and the equity premium puzzle". In: *The Quarterly Journal of Economics* 110.1, pp. 73–92.

— (1999). "Risk aversion or myopia? Choices in repeated gambles and retirement investments". In: *Management science*, pp. 364–381.

— (2007). "Heuristics and biases in retirement savings behavior". In: *The Journal of Economic Perspectives* 21.3, pp. 81–104.

Bernheim, B.D. (2008). "The Psychology and Neurobiology of Judgment and Decision Making: What's in it for Economists?" In: *Neuroeconomics: decision making and the brain*. Ed. by P.W. Glimcher. Academic Press. Chap. 9, pp. 115–125.

Bernoulli, D. (1954). "Exposition of a new theory on the measurement of risk". In: *Econometrica: Journal of the Econometric Society*, pp. 23–36.
Blanchard, D.C. and R.J. Blanchard (1972). "Innate and conditioned reactions to threat in rats with amygdaloid lesions". In: *Journal of Comparative and Physiological Psychology* 81.2, p. 281.
Block, N. (1995). "Some concepts of consciousness". In: *Sciences* 18, p. 2.
Block, N.J., O.J. Flanagan, and G. Güzeldere (1997). *The nature of consciousness: philosophical debates*. MIT Press.
Bonabeau, E. (2002). "Agent-based modelling: Methods and techniques for simulating human systems". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.Suppl 3, p. 7280.
Bower, G.H. (1981). "Mood and memory." In: *American psychologist* 36.2, p. 129.
Bowker, G.C. (2005). *Memory practices in the sciences*. MIT Press.
Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.
— (1999). *Faces of intention: Selected essays on intention and agency*. Cambridge University Press.
Buonomano, D.V. and M.M. Merzenich (1998). "Cortical plasticity: from synapses to maps". In: *Annual review of neuroscience* 21.1, pp. 149–186.
Byrne, A. and M. Tye (2006). "Qualia ain't in the head". In: *Noûs* 40.2, pp. 241–255.
Cahill, L. et al. (1996). "Amygdala activity at encoding correlated with long-term, free recall of emotional information". In: *Proceedings of the National Academy of Sciences* 93.15, pp. 8016–20.
Cahn, B.R. and J. Polich (2006). "Meditation states and traits: EEG, ERP, and neuroimaging studies". In: *Psychological bulletin* 132.2, p. 180.
Cannon, W.B. (1927). "The James-Lange theory of emotions: A critical examination and an alternative theory". In: *The American Journal of Psychology* 39.1/4, pp. 106–124.
Carr, N.G. (2003). "IT doesn't matter". In: *Havard Business Review* 305.
— (2005). "The end of corporate computing". In: *MIT Sloan Management Review* 46.3, pp. 67–73.
Chalmers, D.J. (1996). *The conscious mind: in search of a fundamental theory*. Oxford University Press.
— (2004). "The Representational Character of Experience". In: *The Future for Philosophy*. Ed. by B. Leiter. Oxford University Press.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Vol. 119. The MIT press.
Cohen, P.R. and E.A. Feigenbaum (1982). *The handbook of artificial intelligence*. HeurisTech Press and William Kaufmann.
Cohen, P.R. and H.J. Levesque (1990). "Intention is choice with commitment". In: *Artificial intelligence* 42.2-3, pp. 213–261.
— (1991). "Teamwork". In: *Nous* 25.4, pp. 487–512.
Crittenden, J. (1997). "What Should We Think about Wilber's Method?" In: *Journal of Humanistic Psychology* 37.4, p. 99.
Croft, W.B. (1987). "Approaches to intelligent information retrieval". In: *Information Processing & Management* 23.4, pp. 249–254.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. Jossey-Bass Publishers, San Francisco.
Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon Books, New York.
Damasio, A.R. (1994). *Descartes' error: emotion, reason, and the human brain*. G.P. Putnam.
— (1999). *The feeling of what happens: body and emotion in the making of consciousness*. Harcourt Brace.
Damasio, A.R., BJ Everitt, and D. Bishop (1996). "The somatic marker hypothesis and the possible functions of the prefrontal cortex". In: *Philosophical transactions: Biological sciences*, pp. 1413–1420.
Darwin, C. (1872). *The expression of the emotions in man andanimals*. Oxford University Press, New York.
Dastani, M. and J.J.C. Meyer (2006). "Programming agents with emotions". In: *Proceedings of the 2006 conference on ECAI 2006: 17th European Conference on Artificial Intelligence*. IOS Press, pp. 215–219.
Davidson, D. (2004). *Problems of rationality*. Oxford University Press.
Davidson, R.J., K.R. Scherer, and H.H. Goldsmith (2003). *Handbook of affective sciences*. Oxford University Press, USA.
Dawkins R. Rose, Ch. (Sept. 27, 2005). *A conversation with biologist Richard Dawkins*. `http://www.charlierose.com/view/interview/720/`.
De Martino, B. et al. (2006). "Frames, biases, and rational decision-making in the human brain". In: *Science* 313.5787, p. 684.
Deecke, L., P. Scheid, and H. Kornhuber (1969). "Distribution of readiness potential, pre-motion positivity, and motor potential of the human cerebral cortex preceding voluntary finger movements". In: *Experimental Brain Research* 7 (2), pp. 158–168.
Dennett, D.C. (1988). "Quining qualia". In: *Consciousness in modern science*.
— (1989). *The intentional stance*. The MIT Press.
— (1991). *Consciousness explained*. Little, Brown and Co.
Denton, D.A. (2005). *The primordial emotions: The dawning of consciousness*. Oxford University Press, New York.
Dolan, R.J. (2002). "Emotion, cognition, and behavior". In: *Science* 298.5596, pp. 1191–94.
Dretske, F.I. (1995). *Naturalizing the mind*. MIT Press.
Drucker, P.F. (1969). *The age of discontinuity: Guidelines to our changing society*. Harper and Row New York.
— (1993). *Post-capitalist society*. Harpercollins.
Eagleman, D. (2011). *Incognito: The Secret Lives of the Brain*. Text Publishing Company.
Edgeworth, F. Y. (1881). *Mathematical Psychics*. McMaster University Archive for the History of Economic Thought.
Edwards, W. (1954). "The theory of decision making". In: *Psychological Bulletin* 51.4, pp. 380–417.
— (1961). "Behavioral decision theory". In: *Annual review of psychology* 12.1, pp. 473–498.

Edwards, W., R.F. Miles, and D. Von Winterfeldt (2007). *Advances in decision analysis: from foundations to applications*. Cambridge University Press.

Einstein, A., B. Podolsky, N. Rosen, et al. (1935). "Can quantum-mechanical description of physical reality be considered complete?" In: *Physical review* 47.10, pp. 777–780.

Ekman, P. (1989). "The argument and evidence about universals in facial expressions of emotion". In: *Handbook of social psychophysiology* 58, pp. 342–353.

— (1993). "Facial expression and emotion". In: *American Psychologist* 48.4, p. 384.

Ekman, P., W.V. Friesen, and P. Ellsworth (1982). *Emotion in the human face*. Pergamon Press.

Ellsberg, D. (1961). "Risk, ambiguity, and the Savage axioms". In: *The Quarterly Journal of Economics* 75.4, pp. 643–669.

Emerson, E. and J. Srinivasan (1989). "Branching time temporal logic". In: *Linear Time, Branching Time and Partial Order in Logics and Models for Concurrency*. Ed. by J. de Bakker, W. de Roever, and G. Rozenberg. Vol. 354. Lecture Notes in Computer Science. Springer, pp. 123–172.

Eraker, S.A. and H.C. Sox (1981). "Assessment of patients' preferences for therapeutic outcomes". In: *Medical decision making* 1.1, p. 29.

Espinas, A.V. (1897). *Les Origines de la Technologie*. Alcan, Paris.

Evans, D. (2001). *Emotion: The science of sentiment*. Oxford University Press.

Eysenck, H. J., W. Arnold, and R. Meili, eds. (1972). *Encyclopedia of Psychology*. Search Press, London.

Fischler, I. and G.O. Goodman (1978). "Latency of associative activation in memory." In: *Journal of Experimental Psychology: Human Perception and Performance* 4.3, p. 455.

Fox, C.R. and A. Tversky (1995). "Ambiguity aversion and comparative ignorance". In: *The Quarterly Journal of Economics* 110.3, p. 585.

Freedman, D.A. and R.A. Purves (1969). "Bayes' method for bookies". In: *The Annals of Mathematical Statistics* 40.4, pp. 1177–1186.

Frey, T. (2006). "The future of libraries: Beginning the great transformation". In: *DaVinci Institute, www. davinciinstitute. com/page. php*.

Friedman, M. and L.J. Savage (1948). "The utility analysis of choices involving risk". In: *The Journal of Political Economy* 56.4, pp. 279–304.

Frijda, N.H. (1986). *The emotions*. Cambridge University Press.

Gazzaniga, M.S., ed. (2009). *The cognitive neurosciences*. 4th ed. The MIT Press.

Gershenson, C. (1999). "Modelling emotions with multidimensional logic". In: *Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference of the North American*. IEEE, pp. 42–46.

Gintis, H. (2009). *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton University Press.

Graham, G. (1998). *Philosophy of mind: An introduction*. Vol. 6. Wiley-Blackwell.

Greene, J. (2003). "From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology?" In: *Nature Reviews Neuroscience* 4.10, pp. 846–850.

Greene, J.D. et al. (2004). "The neural bases of cognitive conflict and control in moral judgment". In: *Neuron* 44.2, pp. 389–400.

Greenlaw, R., H.J. Hoover, and W.L. Ruzzo (1995). *Limits to parallel computation: P-completeness theory*. Oxford University Press, USA.

Greenspan, S.I. and S. Shanker (2004). *The first idea: How symbols, language, and intelligence evolved from our early primate ancestors to modern humans*. Da Capo Press.

Gross, A.G. and A.E. Walzer (2008). *Rereading Aristotle's Rhetoric*. Southern Illinois University Press.

Guizzo, E.M. (2003). "The Essential Message: Claude Shannon and the Making of Information Theory". PhD thesis. Massachusetts Institute of Technology.

Hall, R.E. and M. Lieberman (2001). *Economics: principles and applications*. South-Western College Publishing.

Harel, D. and A. Pnueli (1985). "On the Development of Reactive Systems". In: *Logics and models of concurrent systems*, p. 477.

Heath, C. and A. Tversky (1991). "Preference and belief: Ambiguity and competence in choice under uncertainty". In: *Journal of Risk and Uncertainty* 4.1, pp. 5–28.

Heller, M. (2008). *Podglądanie Wszechświata*. Wydawnictwo Znak, Kraków.

Hilgard, E.R. (1980). "The trilogy of mind: Cognition, affection, and conation". In: *Journal of the History of the Behavioral Sciences* 16.2, pp. 107–117.

Hoek, W. Van der, B. Van Linder, and J.J.C.H. Meyer (1999). "An Integrated modal approach to rational agents". In: *Foundations of rational agency*. Ed. by M.J. Wooldridge and A. Rao. Vol. 14. Springer Netherlands, p. 133.

Hoek, W. Van der and M. Wooldridge (2003). "Towards a logic of rational agency". In: *Logic Journal of IGPL* 11.2, pp. 135–159.

Hofstadter, D.R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Basic Books (New York).

Hogarth, R. M. and M. W. Reder (1987). *Rational choice: the contrast between economics and psychology*. University of Chicago Press.

Hume, D. (2003). *A treatise of human nature*. Dover Publications.

Hyman, D.N. (1989). *Modern microeconomics: Analysis and applications*. Irwin.

Ingram, J.K. (1888). "A History of Political Economy". In: *History of Economic Thought Books*.

Ingwersen, P. (1992). *Information retrieval interaction*. s 246. Taylor Graham London.

Jackson, F. (1982). "Epiphenomenal qualia". In: *The Philosophical Quarterly* 32.127, pp. 127–136.

James, W. (1884). "What is an Emotion?" In: *Mind* 9.34, pp. 188–205.

— (1890). *The Principles of Psychology*. Henry Hold and Company.

Jiang, H., J.M. Vidal, and M.N. Huhns (2007). "EBDI: an architecture for emotional agents". In: *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. ACM, p. 11.

Johnson, E.J. and A. Tversky (1983). "Affect, generalization, and the perception of risk". In: *Journal of personality and social psychology* 45.1, pp. 20–31.

Johnson, M.K., M. Verfaellie, and J. Dunlosky (2008). "Introduction to the special section on integrative approaches to source memory." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34.4, p. 727.

Johnson, W.E. (1913). "The pure theory of utility curves". In: *The Economic Journal* 23.92, pp. 483–513.
Kabat-Zinn, J. (1990). *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness*. Delta.
Kaczmarek, J. and D. Ryżko (2009). "Quantifying and optimising user experience: Adapting AI methodologies for Customer Experience Management". In: *Customer Experience Management - Informational Approach to Driving User Centricity*. Ed. by B. Jacobfeuerborn. MOST Press.
Kahneman, D., J.L. Knetsch, and R.H. Thaler (1990). "Experimental tests of the endowment effect and the Coase theorem". In: *The Journal of Political Economy* 98.6, pp. 1325–1348.
— (1991). "Anomalies: The endowment effect, loss aversion, and status quo bias". In: *The Journal of Economic Perspectives* 5.1, pp. 193–206.
Kahneman, D. and A. Tversky (1979). "Prospect theory: An analysis of decision under risk". In: *Econometrica: Journal of the Econometric Society*, pp. 263–291.
— (1984). "Choice, value, and frames". In: *American Psychologist* 39.4, pp. 341–50.
Kandel E. Rose, Ch. (Oct. 29, 2009). *The Great Mysteries of the Human Brain - The Charlie Rose Brain Series*. `http://www.charlierose.com/view/interview/10694?sponsor_id=1/`.
Kandel, E.R. (2006). *In search of memory: The emergence of a new science of mind*. WW Norton & Co.
Kelso, J.A.S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. The MIT Press.
Klein, G.A. (1999). *Sources of power: How people make decisions*. MIT Press.
Knutson, B. et al. (2007). "Neural predictors of purchases". In: *Neuron* 53.1, pp. 147–156.
Knutson, B. et al. (2008). "Neural antecedents of the endowment effect". In: *Neuron* 58.5, pp. 814–822.
Kolb, B. (1995). *Brain plasticity and behavior*. Lawrence Erlbaum Associates.
Korsgaard, C.M. and O. O'Neill (1996). *The sources of normativity*. Cambridge University Press.
Kosslyn, S.M. (1996). *Image and brain: The resolution of the imagery debate*. The MIT Press.
Kotarbiński, T. (1955). *Traktat o dobrej robocie*. Zakład im. Ossolińskich we Wroclawiu.
— (1957). *Myśli o działaniu*. Vol. 1. Państwowe Wydawnictwo Naukowe.
— (1965). *Praxiology: an introduction to the sciences of efficient action*. Pergamon.
Kotarbiński, T. and K. Szaniawski (1972). *Abecadło praktyczności*. Wiedza Powszechna.
Köhler, W. (1929). *Gestalt psychology*. H. Liveright.
Krantz, D.H. (1991). "From indices to mappings: The representational approach to measurement". In: *Frontiers of Mathematical Psychology*, pp. 1–52.
Kripke, S.A. (1980). *Naming and necessity*. Wiley-Blackwell.
Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking Adult.
Lashley, K.S. (1958). "Cerebral organization and behavior." In: *Research publications-Association for Research in Nervous and Mental Disease* 36, p. 1.

Lazarus, R.S. (1990). "Handbook of personality: Theory and research". In: ed. by L.A. Pervin. Guilford Press. Chap. Emotion and adaptation, pp. 609–637.
— (1991). *Emotion and adaptation*. Oxford University Press, USA.
— (2006). *Stress and emotion: A new synthesis*. Springer Publishing Company.
LeDoux, J.E. (1996). *The emotional brain: the mysterious underpinnings of emotional life*. Simon & Schuster.
— (2000). "Emotion circuits in the brain." In: *Annual Review of Neuroscience*.
— (2002). *Synaptic self: how our brains become who we are*. Viking.
Lehar, S. (2006). "The Dimensions of Conscious Experience: A Quantitative Phenomenology". In: *Mind and its place in the world: non-reductionist approaches to the ontology of consciousness*. Ed. by A. Batthyány and A.C. Elitzur. Vol. 7. Ontos Verlag.
Levine, J. (2001). *Purple haze: the puzzle of consciousness*. Oxford University Press.
Lewis, C.I. (1929). *Mind and the world-order: outline of a theory of knowledge*. Charles Scrihner's Sons.
Libet, B. (1999). "Do we have free will?" In: *Journal of Consciousness Studies* 6.8-9, pp. 47–57.
Littlepage, G., W. Robison, and K. Reddington (1997). "Effects of Task Experience and Group Experience on Group Performance, Member Ability, and Recognition of Expertise". In: *Organizational Behavior and Human Decision Processes* 69.2, pp. 133–147.
Lloyd, S. (2002). "Computational capacity of the universe". In: *Physical Review Letters* 88.23, p. 237901.
Locke, J. (1801). *An essay concerning human understanding; with Thoughts on the conduct of the understanding*. Mundell.
Loewenstein, G. and J.S. Lerner (2003). "The role of affect in decision making". In: *Handbook of affective science* 619, p. 642.
Loewenstein, G.F. et al. (2001). "Risk as feelings". In: *Psychological bulletin* 127.2, p. 267.
Lycan, W.G. (1973). "Inverted spectrum". In: *Ratio* 15.315, p. 19.
Lyell, S.C. (1837). *Principles of geology: being an inquiry how far the former changes of the earth's surface are referable to causes now in operation*. Vol. 1. J. Kay, Jun. & Brother.
Machlup, F. (1962). *The production and distribution of knowledge in the United States*. Princeton University Press.
Malt, B.C. and P. Wolff (2010). *Words and the Mind: How words capture human experience*. Oxford University Press, USA.
Mange, D. and M. Tomassini (1998). *Bio-inspired computing machines: Towards novel computational architectures*. PPUR Presses Polytechniques.
Marcel, A.J. (1983). "Conscious and unconscious perception: Experiments on visual masking and word recognition". In: *Cognitive psychology* 15.2, pp. 197–237.
Marsella, S., J. Gratch, and P. Petta (2010). "Blueprint for affective computing: a sourcebook". In: ed. by K.R. Scherer, T. Bänziger, and E. Roesch. Oxford University Press. Chap. Computational models of emotion.
Maslow, A.H. (1971). *The farther reaches of human nature*. Viking Press (New York).

McGaugh, J.L. (2000). "Memory–a century of consolidation". In: *Science* 287.5451, p. 248.
— (2004). "The amygdala modulates the consolidation of memories of emotionally arousing experiences". In: *Annual Review of Neuroscience* 27, pp. 1–28.
Mehrabian, A. and J.A. Russell (1974). *An approach to environmental psychology.*
Merzenich, M.M. et al. (1996). "Temporal processing deficits of language-learning impaired children ameliorated by training". In: *Science* 271.5245, p. 77.
Meyer, J.J., W. van der Hoek, and B. van Linder (1999). "A logical approach to the dynamics of commitments". In: *Artificial Intelligence Preprint Series* 14.
Meyer, J.J.C. (2004). "Reasoning about Emotional Agents". In: *Proceedings of European Conference on Artificial Intelligence – ECAI'04.* IOS Press, pp. 129–133.
Meyer, J.J.C. and W.V.D. Hoek (1995). "Epistemic Logic for AI and Computer Science". In:
Miller, G.A., E. Galanter, and K.H. Pribram (1960). *Plans and the structure of behavior.* Henry Holt and Co.
Millgram, E. (2001). *Varieties of practical reasoning.* MIT Press.
Minsky, M. (1988). *The society of mind.* Simon and Schuster.
Minsky, M.L. (1968). *Semantic information processing.* The MIT Press.
Mises, L. von (1949). *Human action: A treatise on economics.* New Haven: Yale University Press.
Morgan, D.L. (2002). *Essentials of Learning and Cognition.* McGraw Hill.
Muraszkiewicz, M. (June 2011). *Lecture materials on knowledge representation.* URL: `http://www.icie.com.pl/lect_pw.htm`.
Nagel, T. (1974). "What is it like to be a bat?" In: *The Philosophical Review* 83.4, pp. 435–450.
Nahl, D. and D. Bilal (2007). *Information and emotion: the emergent affective paradigm in information behavior research and theory.* Information Today.
Neu, J. (2000). *A tear is an intellectual thing: the meanings of emotion.* Oxford University Press.
Newell, A. and H.A. Simon (1961). "Computer simulation of human thinking". In: *Science* 134.3495, pp. 2011–2017.
Oakley, J. (1992). *Morality and the emotions.* Routledge.
OECD. "Education at a Glance 2011: OECD Indicators". In: URL: `http://dx.doi.org/10.1787/eag-2011-en` (visited on May 6, 2012).
Oliveira, E. and L. Sarmento (2002). "Emotional valence-based mechanisms and agent personality". In: *Advances in Artificial Intelligence*, pp. 771–780.
Ollendick, T.H. (1983). "Reliability and validity of the revised Fear Survey Schedule for Children (FSSC-R)". In: *Behaviour Research and Therapy* 21.6, pp. 685–692.
Ortony, A., G.L. Clore, and A. Collins (1988). *The cognitive structure of emotions.* Cambridge University Press.
Öhman, A., A. Flykt, and F. Esteves (2001). "Emotion drives attention: Detecting the snake in the grass". In: *Journal of Experimental Psychology: General* 130.3, p. 466.

Padgham, L. and G. Taylor (1997). "A system for modelling agents having emotion and personality". In: *Intelligent Agent Systems Theoretical and Practical Issues*, pp. 59–71.

Panksepp, J. (1998). *Affective neuroscience: the foundations of human and animal emotions*. Series in affective science. Oxford University Press.

— (2000). "Emotions as natural kinds within the mammalian brain". In: ed. by M. Lewis and J.M. Haviland-Jones. 2nd ed. Guilford Press New York, pp. 137–156.

— (2005). "Affective consciousness: Core emotional feelings in animals and humans". In: *Consciousness and cognition* 14.1, pp. 30–80.

— (2008). "The affective brain and core consciousness: How does neural activity generate emotional feelings?." In: *Handbook of Emotions, Third Edition*. Ed. by M. Lewis, J.M. Haviland-Jones, and L.F. Barrett. Guilford Publications. Chap. 4, pp. 47–67.

Panksepp, J. and G. Campbell (Jan. 13, 2010). *Brain Science Podcast, Episode #65, Interview with Dr. Jaak Panksepp, Author of Affective Neuroscience: The Foundations of Human and Animal Emotions*. `http://www.brainsciencepodcast.com/bsp/2010/1/13/affective-neuroscience-with-jaak-panksepp-bsp-65.html`.

Pareto, V. (1906). *Manuale di economia politica*. Societa Editrice.

— (1971). *Manual of political economy*. Augustus m Kelley Pubs.

Pawlak, Z. (1981). "Information systems theoretical foundations". In: *Information systems* 6.3, pp. 205–218.

Penfield, W. (1975). *The mystery of the mind: A critical study of consciousness and the human brain*. Princeton University Press.

Pereira, D., E. Oliveira, and N. Moreira (2006). "Modelling emotional BDI agents". In: *Workshop on Formal Approaches to Multi-Agent Systems (FAMAS2006)*.

Persky, J. (1995). "Retrospectives: The ethology of homo economicus". In: *The journal of economic perspectives* 9.2, pp. 221–231.

Picard, R.W. (1997). *Affective computing*. MIT Press.

Pounds, W. F. (1969). "The Process of Problem Finding". In: *Industrial Management Review* 11, pp. 1–19.

Pszczołowski, T. (1967). *Zasady sprawnego działania: wstęp do prakseologii*. Wiedza Powszechna.

Publications, APA and Communications Board Working Group on Journal Article Reporting Standards (2008). "Reporting Standards for Research in Psychology". In: *American Psychologist* 63.9, pp. 839–851.

Rao, A.S. and M.P. Georgeff (1991). "Modeling Rational Agents within a BDI-Architecture". In: *Proceedings of the second International Conference on Principles of Knowledge Representation and Reasoning(KR'91)*. Ed. by J. Allen, R. Fikes, and E. Sandewall. Morgan Kaufmann, pp. 473–484.

Read, D. and G. Loewenstein (1995). "Diversification bias: Explaining the discrepancy in variety seeking between combined and separated choices". In: *Journal of Experimental Psychology: Applied* 1.1, pp. 34–49.

Reisman, G. (1998). *Capitalism: A treatise on economics*. Jameson Books Ottawa.

Reiter, R. (1980). "A logic for default reasoning". In: *Artificial intelligence* 13.1-2, pp. 81–132.

Rescher, N. (2009). "Process Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2009.

Reyna, VF (1995). "Interference effects in memory and reasoning: A Fuzzy-Trace Theory Analysis". In: *Interference and inhibition in cognition*, pp. 29–59.

Rick, S. and G. Loewenstein (2008). "The Role of Emotion in Economic Behavior". In: *Handbook of emotions*. Ed. by M. Lewis, J.M. Haviland-Jones, and L.F. Barrett. 3rd Edition. Guilford Press. Chap. 9, pp. 138–156.

Robbins, L. (1932). *An essay on the nature and significance of economic science*. Macmillan.

Robson, A.J. (1996). "A biological basis for expected and non-expected utility". In: *Journal of economic theory* 68.2, pp. 397–424.

Russell, J.A. (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161.

— (2003). "Core affect and the psychological construction of emotion." In: *Psychological review* 110.1, p. 145.

Russell, J.A. and G. Pratt (1980). "A description of the affective quality attributed to environments". In: *Journal of Personality and Social Psychology* 38.2, p. 311.

Russell, S.J. and P. Norvig (2009). *Artificial intelligence: a modern approach*. Prentice hall.

Rutkowski, L. (2005). *Metody i techniki sztucznej inteligencji: inteligencja obliczeniowa*. Wydawnictwo Naukowe PWN.

Ryle, G. (1949). *The concept of mind*. Barnes & Noble.

Ryżko, D. and J. Kaczmarek (2011). "Customer Experience Management architecture for enhancing corporate customer centric capabilities". Accepted. To be published in ISMIS 2011 Conference Proceedings in Springer's series "Studies in Computational Intelligence".

Salzman, C.D. and W.T. Newsome (1994). "Neural mechanisms for forming a perceptual decision". In: *Science* 264.5156, p. 231.

Samuelson, P.A. (1947). *Foundations of economic analysis*. Harvard University Press, Cambridge, Mass.

Sapolsky, R.M. (1996). "Why stress is bad for your brain". In: *Science* 273.5276, p. 749.

Savage, L.J. (1954). *Foundations of statistics*. Wiley.

Schacter, D.L. (2002). *The seven sins of memory: How the mind forgets and remembers*. Mariner Books.

Schacter, D.L. and D.R. Addis (2007). "The cognitive neuroscience of constructive memory: Remembering the past and imagining the future". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481, p. 773.

Scherer, K.R., T. Bänziger, and E. Roesch (2010). *Blueprint for affective computing: a sourcebook*. Oxford University Press.

Scherer, K.R., A. Schorr, and T. Johnstone (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Schmitt, B. (2003). *Customer experience management: a revolutionary approach to connecting with your customers*. John Wiley & Sons Inc.

Schwartz, J. and S. Begley (2002). *The mind and the brain: neuroplasticity and the power of mental force*. Regan Books.

Schwartz, J.M. and S. Begley (2003). *The mind and the brain: Neuroplasticity and the power of mental force*. Harper Perennial.
Searle, J. R. (1999). *Mind, Language, and Society : Philosophy in the Real World*. Basic Books.
Searle, J.R. (1982). "The myth of the computer". In: *The New York Review of Books* 29.7, pp. 3–6.
— (1983). *Intentionality, an essay in the philosophy of mind*. Cambridge University Press.
— (2001). *Rationality in action*. MIT Press.
— (2002). *Consciousness and language*. Cambridge Univ Pr.
— (2004). *Mind: a brief introduction*. Oxford University Press.
— (2008). *Freedom and neurobiology: Reflections on free will, language, and political power*. Columbia University Press.
Searle, J.R., D.C. Dennett, and D.J. Chalmers (1997). *The mystery of consciousness*. New York Review of Books.
Segerberg, Krister, John-Jules Meyer, and Marcus Kracht (2009). "The Logic of Action". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2009. Standford University.
Shafir, E. and R.A. LeBoeuf (2002). "Rationality". In: *Annual Review of Psychology* 53.1, pp. 491–517.
Shannon, C.E. (1938). "A Symbolic Analysis of Relay and Switching Circuits". In: *Transactions of the American Institute of Electrical Engineers* 57.12, pp. 713–723.
— (2001). "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1, pp. 3–55.
Shapiro Jr, D.H. and R.N. Walsh (2008). *Meditation: Classic and contemporary perspectives*. Aldine De Gruyter.
Shaw, C.A., J.C. McEachern, and J. McEachern (2001). *Toward a theory of neuroplasticity*. Psychology Press. ISBN: 9781841690216. URL: `http://www.google.com/books?id=cbeY\_fm0YPgC`.
Siewert, C.P. (1998). *The significance of consciousness*. Princeton University Press.
Simon, H.A. (1959). "Theories of decision-making in economics and behavioral science". In: *The American Economic Review* 49.3, pp. 253–283.
— (1967). "Motivational and emotional controls of cognition". In: *Psychological review* 74.1, pp. 29–39.
— (1973). "The structure of ill structured problems". In: *Artificial intelligence* 4.3-4, pp. 181–201.
— (1978). "Information-Processing Theory of Human Problem Solving". In: *Human information processing* 5, p. 271.
Simon, H.A. and A. Newell (1958). "Heuristic problem solving: The next advance in operations research". In: *Operations research* 6.1, pp. 1–10.
Sloman, A. (1987). "Motives Mechanisms Emotions". In: *Emotion and Cognition* 1.3, pp. 217–234.
— (1999). "What sort of architecture is required for a human-like agent". In: *Foundations of Rational Agency*, pp. 35–52.

Sloman, A. and M. Croucher (1981). "Why robots will have emotions". In: *Proceedings of the 7th international joint conference on Artificial intelligence*. Vol. 1. Morgan Kaufmann Publishers Inc., pp. 197–202.

Sloterdijk, P. (1987). *Critique of cynical reason*. University of Minnesota Press.

Slovic, P., B. Fischhoff, and S. Lichtenstein (1982). "Facts versus fears: Understanding perceived risk". In: *Judgment under uncertainty: Heuristics and biases*. Ed. by D. Kahneman, P. Slovic, and A. Tversky. Cambridge University Press, pp. 463–489.

Slutsky, E. (1915). "Sulla teoria del bilancio del consumatore". In: *Giornale degli economisti* 51, pp. 1–26.

Smith, H. and P. Fingar (2003). *IT doesn't matter – business processes do: a critical analysis of Nicholas Carr's IT article in the Harvard business review*. Meghan-Kiffer Press.

Smith, J.M. and G.R. Price (1973). "The Logic of Animal Conflict". In: *Nature* 246.5427, pp. 15–18.

Solomon, R.C. (1973). "Emotions and choice". In: *The Review of Metaphysics*, pp. 20–41.

Soon, C.S. et al. (2008). "Unconscious determinants of free decisions in the human brain". In: *Nature neuroscience* 11.5, pp. 543–546.

Sosińska-Kalata, B. (1999). *Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach*. Stowarzyszenie Bibliotekarzy Polskich.

Stein, N.L., M.W. Hernandez, and T. Trabasso (2008). "Advances in modeling emotion and thought: The importance of developmental online, and multilevel analyses". In: *Handbook of Emotions, Third Edition*. Ed. by M. Lewis, J.M. Haviland-Jones, and L.F. Barrett. Guilford Publications. Chap. 35, pp. 574–586.

Steunebrink, B.R., M. Dastani, and J.J.C. Meyer (2007). "A logic of emotions for intelligent agents". In: *Proceedings of the 22nd national conference on Artificial intelligence*. Vol. 1. AAAI Press, pp. 142–147.

Steup, Matthias (2008). "The Analysis of Knowledge". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2008.

Stone, A., T. Valentine, and R. Davis (2001). "Face recognition and emotional valence: Processing without awareness by neurologically intact participants does not simulate covert recognition in prosopagnosia". In: *Cognitive, Affective, & Behavioral Neuroscience* 1.2, pp. 183–191.

Strawson, G. (1994). *Mental reality*. The MIT Press.

Tarski, A. (1933). *Pojęcie prawdy w językach nauk dedukcyjnych*. Towarzystwo Naukowe Warszawskie, Wydz.III.

Thaler, R.H. et al. (1997). "The Effect of Myopia and Loss Aversion on Risk Taking: An Experimental Test". In: *The Quarterly Journal of Economics* 112.2, pp. 647–661.

Thomson, J.J. (1985). "The Trolley Problem". In: *Yale Law Journal* 94.6, pp. 1395–1415.

Tinbergen, N. (1951). "The study of instinct." In:

Toffler, A. (1984). *Future shock*. Bantam.

Tom, S.M. et al. (2007). "The neural basis of loss aversion in decision-making under risk". In: *Science* 315.5811, p. 515.

Tulving, E. (1985). "Memory and consciousness". In: *Canadian Psychology/Psychologie Canadienne* 26.1, p. 1.
Turing, A.M. (1936). "On computable numbers, with an application to the Entscheidungs problem". In: *Proceedings of the London Mathematical Society*, p. 42.
— (1950). "Computing machinery and intelligence". In: *Mind* 59.236, pp. 433–460.
Tversky, A. and D. Kahneman (1992). "Advances in prospect theory: Cumulative representation of uncertainty". In: *Journal of Risk and uncertainty* 5.4, pp. 297–323.
Tye, M. *Ten problems of consciousness: a representational theory of the phenomenal mind*. MIT Press.
— (2000). *Consciousness, color, and content*. MIT Press.
— (2003). *Consciousness and persons: Unity and identity*. MIT Press.
Tye, Michael (2009). "Qualia". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2009.
Van Gulick, R. (2011). "Consciousness". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2011.
Van Harmelen, F., V. Lifschitz, and B. Porter (2008). *Handbook of knowledge representation*. Elsevier.
Van Linder, B., W. van der Hoek, and J. Meyer (1995). "Actions that make you change your mind". In: *KI-95: Advances in Artificial Intelligence*, pp. 185–196.
Van Linder, B., J.J.C. Meyer, and W. Van Der Hoek (1997). "Formalizing motivational attitudes of agents using the KARO framework". In: *UU-CS* 1997-03.
Von Neumann, J. and O. Morgenstern (1944). "Theory of games and economic behavior." In:
Wallace, R. Jay (2009). "Practical Reason". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2009.
Weiss, G. (2000). *Multiagent systems: a modern approach to distributed artificial intelligence*. The MIT press.
Whitehead, A.N., D.R. Griffin, and D.W. Sherburne (1979). *Process and reality: An essay in cosmology*. Free Pr.
Wilber, K. (1977a). *Ken Wilber Stops His Brain Waves - an experiment with EEG*. `http://www.youtube.com/watch?v=LFFMtq5g8N4`. (Visited on Jan. 13, 2010).
— (1977b). *The spectrum of consciousness*. Theosophical Publication House.
— (2001). *The eye of spirit: An integral vision for a world gone slightly mad*. Shambhala Boston.
Wilson, R.A. and F.C. Keil (2001). *The MIT encyclopedia of the cognitive sciences*. The MIT Press.
Winkielman, P. et al. (2007). "Affective influence on judgments and decisions: Moving towards core mechanisms". In: *Review of General Psychology* 11.2, p. 179.
Wittgenstein, L. (1958). *Philosophical investigations*. Macmillan.
Wold, H., G.L.S. Shackle, and L.J. Savage (1952). "Ordinal preferences or cardinal utility?" In: *Econometrica* 20.4, pp. 661–664.
Wooldridge, M. and N.R. Jennings (1995). "Intelligent agents: Theory and practice". In: *The knowledge engineering review* 10.02, pp. 115–152.
Wooldridge, M.J. (2000). *Reasoning about rational agents*. The MIT Press.

Wozniak, P.A. (1995). "Economics of learning". PhD thesis. University of Economics, Wroclaw, Poland.

Wozniak, P.A. and E.J. Gorzelanczyk (1994). "Optimization of repetition spacing in the practice of learning". In: *Acta Neurobiologiae Experimentalis* 54, pp. 59–59.

Zajonc, R.B. (1980). "Feeling and thinking: Preferences need no inferences." In: *American psychologist* 35.2, p. 151.

— (1984). "On the Primacy of Affect". In: *American Psychologist* 39.2, pp. 117–23.

Zeleny, M. (2002). "Knowledge of enterprise: knowledge management or knowledge technology?" In: *International Journal of Information Technology and Decision Making* 1.2, pp. 181–207.

Zins, C. (2007). "Conceptions of information science". In: *Journal of the American Society for Information Science and Technology* 58.3, pp. 335–350.

Zweig, J. (Jan. 1998). "How the Big Brains Invest at TIAA-CREF". In: *Money* 27.1, p. 114.

affect. URL: `http://www.merriam-webster.com/dictionary/affect/` (visited on Nov. 8, 2011).

experience. URL: `http://www.britannica.com/bps/search?query=experience` (visited on Feb. 4, 2009).

knowledge. URL: `http://www.merriam-webster.com/dictionary/knowledge/` (visited on Nov. 8, 2011).