5-2016

Modeling human-like non-rationality for social agents

Jaroslaw KOCHANOWICZ

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Daniel THALMANN

Modeling human-like non-rationality for social agents

Jaroslaw Kochanowicz
IMI, SCE, NTU
Singapore 637553
jaroslaw@ntu.edu.sg

Ah-Hwee Tan
SCE, NTU
Singapore, 639798
Dublin, Ohio 43017-6221
asahtan@ntu.edu.sg

Daniel Thalmann
IMI, NTU
Singapore, 637553
danielthalmann@ntu.edu.sg

ABSTRACT

Humans are not rational beings. Deviations from rationality in human thinking are currently well documented [25] as non-reducible to rational pursuit of egoistic benefit or its occasional distortion with temporary emotional excitation, as it is often assumed. This occurs not only outside conceptual reasoning or rational goal realization but also subconsciously and often in certainty that they did not and could not take place 'in my case'. Non-rationality can no longer be perceived as a rare affective abnormality in otherwise rational thinking, but as a systemic, permanent quality, 'a design feature' of human cognition. While social psychology has systematically addressed non-rationality of human cognition (including its non-emotional aspects) for decades [63]. It is not the case for computer science, despite obvious relevance for individual and group behavior modeling. This paper proposes brief survey of work in computational disciplines related to human-like non-rationality modeling including: Social Signal Processing, Cognitive Architectures, Affective Computing, Human-Like Agents and Normative Multi-agent Systems. It attempts to establish a common terminology and conceptual frame for this extremely interdisciplinary issue, reveal assumptions about non-rationality underlying the discussed models and disciplines, their current limitations and potential in contributing to solution. Finally, it also presents ideas concerning possible directions of development, hopefully contributing to solution of this challenging issue.

CCS Concepts

•**Computing methodologies** → **Cognitive science;** Modeling methodologies; •**Human-centered computing** → *Social engineering (social sciences);*

Keywords

rationality; irrationality; affective computing;cognitive architectures; social simulation; multi-agent systems

CASA 2016, May 23 - 25, 2016, ,

ISBN 978-1-4503-4745-7/16/05. . . $15.00

DOI: http://dx.doi.org/10.1145/2915926.2915951

1. INTRODUCTION

This paper proposes a survey of work in computational disciplines related to human-like non-rationality modeling including: Social Signal Processing, Cognitive Architectures, Affective Computing, Human-Like Agents and Normative Multi-agent Systems. It attempts to establish a common terminology and conceptual frame for this extremely interdisciplinary issue, to reveal assumptions about non-rationality underlying the discussed models and disciplines, their current limitations and potential in contributing to the solution. Finally, it also presents ideas concerning possible directions of development, hopefully contributing to a solution of this challenging issue.

1.1 Humans are not rational

Humans are not rational beings. Image of eyes above coffee contribution box in the office inclines people to pay more often without supervision [6], without changing the situation from an egoistic-utilitarian or even moral perspective. Attribution of positive feature like intelligence to a person primes us to attribute them with other positive qualities like honesty or generosity, regardless of any statistical correlation they may have, in a 'halo effect' example [62]. Countless examples of this type occur not only outside conceptual reasoning or rational goal realization but also subconsciously and often in certainty that they do not and could not take place 'in my case'. Habitual, gregarious, ritualistic, unconscious, stereotypic and erroneous deviations from rationality in human thinking and behavior are currently well documented [25] as non-reducible to rational pursuit of egoistic benefit or its occasional distortion with temporary emotional excitation, as it is often assumed. Non-rationality can no longer be perceived as a rare affective abnormality in otherwise rational thinking, but as a systemic, permanent quality, 'a design feature' of human cognition.

1.2 Social agents, culture and non-rationality

Tightly related to human non-rationality is the second main topic of this paper - the struggle to understand social and cultural phenomena and its emergence in the micro and macro scales. Traditionally a challenge of social sciences, this problem has growing relevance for Cognitive Architectures (**CA**), Agent Based Systems (**ABS**), Affective Computing (**AF**), Social Signal Processing(**SSP**) and other computational disciplines.

The interaction of human rationality and non-rationality (affect, personality, habit and others) results in the emergence of rituals, customs, stereotypes or prejudice, shaping social process and structure. Subconscious and sub-symbolic aspects of identification, relations, values, power distribution, roles or norms determine objective group character through agent actions as strong as (some would say stronger than) their formal, explicit or rational counterparts, in a process of *culture emergence*.

1.3 Situation in computational models

While social psychology has systematically addressed non-rationality of human cognition (including its non-emotional aspects) for decades [63]. It is not the case for computer science, despite obvious relevance for individual and group behavior modeling. Especially modeling of socio-cultural non-rationality is still a relatively underdeveloped process, despite its great relevance and normative-emotional power. It remains a grand challenge due to its complex, informal and fuzzy nature.

1.4 Structure of the paper

The rest of the paper is organized as follows. In the next section wider background of the problem is presented, including discussion and formalization of the used terms and brief history or rationalistic perspective in social sciences and computer science. In section 3 a survey of work in computational disciplines related to human-like non-rationality modeling is proposed. It includes: Social Signal Processing, Cognitive Architectures, Affective Computing, Human-Like Agents and Normative Multi-agent Systems. Next, possible research directions are discussed. The paper ends with summary.

2. BACKGROUND

To properly frame and present the problem of non-rationality both well-defined terminology and a historical perspective on human behavior modeling are useful and this section provides them shortly.

2.1 Terms and definitions

In this section most relevant terms are defined, to allow a precise presentation. This includes comments on the distinction between rationality, irrationality and non-rationality, definition of 'System 1' and 'System 2' and finally, 'cognitive' vs. 'non-cognitive' process dichotomy.

2.1.1 Rationality, irrationality and non-rationality

In general, *rational* decision making is an analytic process that operates on a formal conceptual model of the situation and goals to be achieved in it. Potential alternative actions are represented similarly and their consequences are calculated in a chain of reasoning that can be described, analyzed and communicated. Evaluation of their expected outcomes from the position of goals allows to select the optimal action.

Irrational decision is one that deviates from the optimal action, regardless of the nature of the decision making process. *Non-rational* decision making on the other hand is one that is not based on the explicit representations, formal reasoning and sequential analysis. It is heuristic and associative and often fast . Also, such process cannot normally be communicated due to the aforementioned lack of formal and explicit representations (making definition of non-rationality dependent on decision what is considered formal/explicit in a given case).

Irrationality therefore deviates from rationality by result, while non-rationality by mode of work and one does not imply the other (although they are often correlated). Example: good heuristic may often provide a non-rational but correct (so not irrational) choice, while error in formal reasoning provides an irrational but not non-rational answer. Although irrationality is sometimes mentioned, the main interest of this paper is (human-like) non-rationality and the optimality of actions is secondary. Notably, this includes in the scope of the paper human-like aspects, whose optimality and therefore irrationality is at least practically problematic to establish e.g. it is hard to say which types of personality are 'irrational' to have, but personality certainly belongs to non-rational aspects of human-likeness.

This last difference between 'non-rational' and 'irrational' becomes crucial, when the goal or motivational aspect of the decision making system is considered. As irrationality is defined based on the optimality of goal realization, it is impossible to describe any goal as irrational, including informal and temporary urges or desires, even harmful or destructive ones. While formalism assessing (at least theoretically) irrationality of specific goals/desires (for example of suicidal/murderous vs. altruistic ones) is probably possible in the specific case of humans (e.g. based on the overall subjective happiness obtained when specific goals are present) human motivation particularities are trivially included as human-like non-rationality.

2.1.2 'System 1' and 'System 2'

In the case of human decision makers, relation between rationality and non-rationality is perhaps best addressed by the dual process theory [63]. Its foundations are as old as modern psychology and likely come from William James who claimed there are two different kinds of thinking: associative and true reasoning. Over the decades, this concept has been developed and reformulated several times and has recently been popularized by Kahneman who uses terms *'System 1'* (non-rational, associative intuition) and *'System 2'* (rational, conceptual reasoning, see Table 1 for details). Dozens of cognitive biases related to 'System 1' are known [25], including :

- 'Anchoring effect' - tendency to be influenced by irrelevant data. For example, when asked if Gandhi was more than 114 years old when he died, people will provide a much larger estimate of his age at death than when anchoring question referred to death at 35
- 'halo effect' - tendency to 'transmit' assessments of others from one area of perceptions to another
- 'framing' - drawing different conclusions from the same information, depending on presentation
- 'pervasive optimistic bias' - tendency to overestimate benefits and underestimate costs e.g. in 2002 American kitchen remodeling was expected on average to cost $18,658, but actually cost $38,769.
- 'substitution' - 'System 1' tendency to substitute a difficult question with a simpler one
- anthropomorphism - tendency to characterize animals, objects, and abstract concepts as possessing human-like traits, emotions, and intentions

Table 1: 'System 1' vs. 'System 2' comparison

'System 1'	'System 2'
Unconscious	Conscious
Implicit	Explicit
Automatic	Controlled
Low Effort	High Effort
Rapid	Slow
Associative	Rule-Based
Contextualized	Abstract
Recognition, perception	Rules, comparisons, analysis
Parallel	Serial
Large Capacity	Small Capacity

2.1.3 *Cognitive and non-cognitive processes*

Another popular distinction relevant to this discussion is the one between cognitive and non-cognitive processes, as 'non-cognitive' is often used where 'non-rational' could perhaps be a better alternative. For example, classification of affect theories into cognitive and non-cognitive ones (see 3.3.1) distinguishes the latter as involving 'only nonpropositional and possibly even nonconceptual mental representations' [56] and the former as requiring 'certain higher-order mental representations, in particular beliefs and desires' [56].

This is somewhat counterintuitive, as 'cognition' means simply *acquisition of knowledge* 'through thought, experience, and the sense' [16] and could perhaps be related to either 'System 1' or 'System 2' acquisition. Indeed, 'usefulness of this distinction has been questioned because of the vagueness and ambiguity of the term cognitive' [56]. Defining 'cognitive' process as acquiring some (even implicit) knowledge within the system (e.g. by drilling basketball throws) and 'non-cognitive' process as free of such acquisition (e.g. the same throws performed by a non-learning robot) while using 'rational vs. non-rational' or 'System 1 vs. System 2' spectrum to describe the mode of this process could perhaps help avoiding some confusion.

However, even if the initial definition is used, clear difference between 'rational' and 'cognitive' processes must be observed, as the latter one simply involves *some* 'higher order' representations or analysis while the former has demands on all steps of the process, and so e.g. cognitive emotion elicitation would still not be a rational process.

2.2 Abandoning homo-economicus

Decision making models in various fields including economy and computational behavior modeling assumed *'homo economicus'* [26] i.e. perspective on humans as basically rational, self-interested and pursuing individual and autonomously selected goals as optimally as possible. To understand the situation in human non-rationality modeling, a certain historical perspective on this concept is needed. Two relevant trends are briefly presented in this section: one in social sciences and another in computer simulation.

2.2.1 *In traditional social sciences*

First trend is a broad transition from rationalistic perspective on humans in social sciences, to one including consistently irrational aspects. While psychologists always emphasized profound influence of strong emotion on human behavior, core of the new perspective is that source of human irrationality is also non-affective. The new message is: we are irrational even without emotional excitation. As Kahneman puts it 'at any given time scholars in a particular field tend to share basic assumptions about their subject. Social scientists in the 1970s broadly accepted two ideas about human nature. First, people are generally rational, and their thinking is normally sound. Second, emotions such as fear, affection, and hatred explain most of the occasions on which people depart from rationality. Our article challenged both assumptions.(...) We documented systematic errors in the thinking of normal people, and we traced these errors to the design of the machinery of cognition rather than to the corruption of thought by emotion' [28]. This paradigm shift cannot be overestimated in any context of behavior modeling - a fact reflected by Kahneman's Nobel Prize in Economy.

2.2.2 *In computer sciences*

Second trend is a gradual inclusion of human-like elements in Cognitive Architectures, Social Simulation, Multi Agents Systems etc. While its traces may be tracked to the early AI, mainstream AI embraced those notions only in the early 90's, when fields like Affective Computing or now-classical emotion models like OCC [51] gained attention. Only relatively recently non-rationality became widely addressed in the computational science and until now mostly in its most traditional i.e. affective dimension. This delay is understandable due to the engagement in pragmatic 'rational' challenges (planning, image recognition, language processing etc.), background of the researchers and what Picard called emotion's 'stigma' in science [54]. Moreover, non-affective non-rationality modeling suffers from inherent inaccessibility of subconscious structures and limited data for validation of the models. Following analysis of how most relevant domains and specific models include aspects of non-rationality, will provide opportunity for concrete examples.

3. RELEVANT DOMAINS

This section discusses how achievements and limitations of existing domains are linked to the problem of human-like non-rationality, especially in the social context. The addressed fields are: Social Signal Processing, Cognitive Architectures, Affective Computing, Social Agents with personality and Normative ABS in Computational Sociology.

3.1 Social Signal Processing

Growing domain of Social Signal Processing (**SSP**) is interested in imitating a set of human 'ability to express and recognize social signals like turn taking, agreement, politeness, empathy, friendliness, conflict, etc.' [52] and to 'manage them in order to get along well with others while winning their cooperation' [52]. The ultimate goal of SSP are computer systems 'capable of sensing agreement, inattention, or dispute, and capable of adapting and responding in real-time to these social signals in a polite, non-intrusive, or persuasive manner, are likely to be perceived as more natural, efficacious, and trustworthy' [52]. Three tasks are therefore at its core, namely analysis, modeling and generation of social signals, understood as communicative and informative messages (conscious or not) related to social facts. These social facts are typically categorized as:

- social interaction - specific instance of social event signaled through nods manifesting agreement, blink manifesting intimacy etc.

- social emotion - subset of emotion related to others like envy or empathy, signaled by tone, gestures etc.
- social evaluation - assessments of characteristics of a person or group based on certain standards.
- social attitudes - a broad category defined as a tendency to behave in a certain way towards person or a group.
- social relation - relation towards agent(s) or group(s) based on co-dependence of their goals e.g. dependency, competition, exploitation etc. This concept notably includes social roles understood as behavior expectation patterns.

Signals related to the above social facts manifest through multi-modal cues like 'facial expressions, body postures and gestures, vocal outbursts like laughter, etc.' which can be automatically analyzed by technologies of signal processing. This domain is at its early stage, it is still addressing its basic research questions and many problems have only been addressed recently. Works addressing non-verbal social interaction analysis in small groups have been presented by Gatica-Perez [23], Gunes surveys recognition of emotion [24] and Kleinsmith affective bodily expression recognition [32]. Work on evaluations and attitudes, is more scarce. Examples include facial attractiveness estimation [27], automatic assessment of agreement [27]. In comparison, there is a great number of works related to social relations (roles in particular) e.g. recognition of roles in shows [57] and meetings [53].

3.1.1 Relation to non-rationality and limitations

From the perspective of SSP, human-like non-rationality should not be seen as a separate category but rather one pervading all the mentioned social fact classes and cognitive processes causing them. For example, social evaluations and attitudes have been proven to be heavily influenced by specific 'System 1' particularities, like the mentioned cognitive biases. Social relations often depend on non-rational, context and culture dependent patterns that have no justification but a habitually repeated cultural transmission. All of them are known to rely on stereotypes maintained to support mentioned habits or mental processes like self-identification and other human-specific social needs.

This relates to a potential weakness of SSP as an approach to human-like non-rational phenomena modeling. The very term 'Social Signal' ties attention to the *surface, specific and behavioral* aspect i.e. visible communication, while developments in cognitive science and psychology provided knowledge about *general, deep and cognitive* characteristics and limitations that influence not only social interaction. Popular solutions often used in SSP like hidden Markov models, conditional random fields or deep belief networks, are pragmatic and effective in specific tasks, but they do not take advantage of knowledge about human 'System 1' nor help to model it in a broader scope. In this, we agree with Pantic et. al. who pointed out that SSP would benefit from addressing some of its challenges, specifically signal contextuality as 'one complex problem rather than a number of detached problems in human sensing, context sensing, and human behaviour understanding'[52]. Therefore, signal processing techniques effective at particular tasks (like gesture recognition) need to be placed in a context of general cognitive architectures (discussed in the next section) to provide a broad computational theory of human-like non-rationality.

3.2 Cognitive Architectures

Requirement to place a specified theory in a concrete, 'default' cognitive environment (as postulated above) is a typical problem faced in application or validation in various domains (including SSP) and one of the motivations to develop Cognitive Architectures. Cognitive Architecture (**CA**) is a model of artificial, computational processes, aiming to resemble human cognitive abilities. It possesses a broad spectrum of intelligence, as opposed to being a solution to one particular narrow task. As Ng states 'cognitive architecture can be defined as a single system that is capable of producing all aspects of behavior, while remaining constant across various domains and knowledge bases' [50]. Sun puts it shortly: 'cognitive architecture is a domain-generic computational cognitive model' [60].

When describing CA, it is meaningful to distinguish between psychologically and engineering oriented ones. The latter are designed for purely pragmatically purposes, while the former aim at a relative structural and functional cognitive realism and human-likeness, and are designed as a research tool of cognitive scientists: 'Cognitive architecture provides a concrete framework for more detailed modeling of cognitive phenomena, through specifying essential structures, divisions of modules, relations between modules, and so on. Its function is to provide an essential framework to facilitate more detailed modeling and understanding of various components and processes of the mind. It forces one to think in terms of process, and in terms of detail. Instead of using vague, purely conceptual theories, cognitive architectures force theoreticians to think clearly. They are critical tools in the study of the mind' [61]. For the engineering oriented architectures, psychological realism is a minor concern and their design structure reflects problem space, rather than human mind. Besides theoretical importance both psychological and engineering oriented CA also have a direct, practical and commercial use and have been used successfully, to solve complex tasks. For our purposes, psychologically believable ones are more interesting. Examples of prominent CA that are symbolic, production based systems: SOAR [36], Icarus [37] for hybrid neural-symbolic systems examples are Epic [31], Clarion[60] and ACT-R [2]. Many good and detailed overviews exist [13, 18, 38].

The most direct limitation of pure CA from the perspective of human-like non-rationality is the fact that most of them are focused on solving traditional challenges of AI like reasoning, planing, language processing etc. leading to initial marginalization of non-rationality even in its most obvious i.e. affective form. This remark is however not true for Cognitive-Affective Architectures described below.

3.3 Affective Computing, Cognitive Affective Architectures and Human-Like Agents

As mentioned, traditional AI (including traditional CA) has been developed for a long time with a wide disregard to human-like non-rationality, including emotion. Affective Computing (**AC**) is a developing, interdisciplinary discipline that started after Picard's paper [54], as a reaction to this tendency. It encompasses the development of systems that can recognize, model and generate human affects and draws from computer science, psychology, and cognitive science. Besides the ability to understand, react to and generate affect, AC addresses question how affect sensing and generating technology influences human-computer interactions and how it may improve them.

3.3.1 Traditional theories of emotion

Over the decades psychology developed countless incompatible emotion theories, none of them formulated in a computable language, very few formal enough to be a direct base for computational model and most expressed in a way that makes it difficult to relate them to others. This historical heritage creates a monumental task for both computer scientists and modern psychologists attempting to reconstruct those theories in a common computational platform for research, validation or practical application.

There are various ways in which countless theories of emotion can be divided, with perhaps the most popular distinction of emotion elicitation theories into cognitive versus non-cognitive ones (as understood e.g. by Reisenzein et. al. [56], see 2.3.1). In psychology, discussion about the scope of non-cognitive emotion is controversial, partly due to obscurity of the term 'cognitive'. Most relevant cognitive theory (and emotion theory in general) is the appraisal theory [4, 22], 'a predominant force among psychological perspectives on emotion and arguably the most fruitful source for those interested in the design of symbolic AI systems, as it emphasizes and explains the connection between emotion and cognition' [42].

3.3.2 Computational models of emotion

To achieve computational models of emotions, first reconstruction into formal but implementation-independent languages is required. Its goal to is improve the original theories by achieving greater formalization, forcing clarification and disambiguation in preparation for implementation, while avoiding complexities related to implementational details. First language proposed for formalizing psychological emotion theories was the set theory [10] due to its formalism, expressiveness and close relation to computational theories. In this way, general model of cognitive-appraisal theories was initially formalized [10].

Later other formalisms were used, notably representation of emotion theories in agent logics that exchanged some expressiveness for easier specification. Especially belief-desire-intention (**BDI**) model [55] provided a convenient conceptual environment, as most emotion theories use beliefs and desires as their base. The fact that they are already provided by BDI, has allowed an easy specification and on the other hand a link to rational decision making already specified with BDI . Therefore, many theories were realized using BDI e.g. emotions related to expectations [12], four basic emotions [46] and most notably two formalizations [56, 1] of the 22 emotions of the OCC theory [51] that became a dominant standard in many practical applications.

3.3.3 Cognitive-Affective Architectures

Finally, the general purpose CA (see section 3.2) were also used as means of specification of affect theories. They offered freedom from irrelevant implementation efforts and decisions and more complete environment for the tested theories by providing the 'default', approximated answers. As a result, a subgroup of CA appeared: Cognitive-Affective Architectures (**CAA**) [65]. CAA addresses not only specification of affect generation model but also relation between affect and rationality, allowing formal and more complete specification of emotion effect theories. CAA have been developed both as tools in academic research, human-machine interfaces [30], entertainment [59], training or advertising [65]. Generally, academic CAA are theoretically more sound, while non-academic CAA address greater spectrum of features. Believable affective modeling of 'Artificial Humans' is a field in itself. Examples of CAA include: Soar-Emote [40], that combined implementation of both physiological and emotion cognitive emotion theories in SOAR or appraisal theory implementation in a computational model EMA [41]. For comprehensive overview of CAA please refer to [39, 65].

3.3.4 Issues with personality in Human-Like Agents

As the development of CAA progressed, virtual [64] and robotic human-like agents made significant advance. For this class of CAA, believability and attractiveness for the users is often the key goal. This caused inclusion of other aspects of human-like non-rationality, notably personality.Concept of personality embodies the assumption about individual behavioral coherence and consistency existing as object-like entity invariable throughout contexts. A century of intensive efforts in identification of this entity proved this assumption surprisingly unfruitful [47], as dispositional differences between individuals are consistently found to be smaller than between-situation differences for one individual [20] (Figure 1). This fact cannot be explained by the prominent 'trait models', like Five Feature Model (FFM)[44] that emerged from this perspective.

Domination of Five Feature Model (FFM) as a personality model in simulations, while understandable due to simple form and easily obtainable parametrization, comes at a price of serious limitations implied by any 'trait model'. Their drawbacks as descriptive tools are multiplied when they are used as generating mechanisms. The fact that 'trait models' are descriptive and are in no sense prescriptive in terms of cognition is vastly ignored in computational modeling. When used for behavior generation those 'average attitude values' produce uniform, flat personality that undermines psychological believability. This is exactly due to the negation of contextual cognitive-affective changes. FFM has been theoretically criticized in psychology, as addressing 'surface, easily noticeable aspects of personality and neglecting more private or *context-dependent ones*' [43].

In practice, 'agreeable' or 'extrovert' FFM based character is 'agreeable' or 'extrovert' in all contexts, with no structured individualized variability, resulting in a predictability that fails to surprise in a new context like a fearful mother's, bravery when her children are threatened. This flaw is not limited to FFM and extends to all 'trait models', like PEN [21] and other, less popular ones. All of this has understandable consequences for individual modeling but, less obviously, models of non-rational components (like personality) are becoming a bottle neck in multi-agent and agent based social simulations [33, 29].

3.3.5 CAA and irrationality

The development of CAA within CA exemplifies the shift towards models including affect and in the broader perspective abandonment of rationalistic paradigm by inclusion of non-rational aspects.

However, even CAA are a) limited to affective non-rationality leaving out other 'System 1' particularities (see 2.1.2) and b) designed with a focus on individual and used in separation (with the exception of Clarion [60]) and ignore many mechanisms of social cognition crucial in sociological psychology. Therefore CAA (including Clarion) do not explicitly model 'System 1' mechanisms, particularly social and cultural ones like social context cognition, stereotypical agent perception or human-like identity management. As a result, interaction of multiple CAA does not manifest rich and complex social structures, and other phenomena found in animals, let alone humans.

CA evolved to CAA by including affect, but have yet to embrace the non-affective and socio-cognitive 'System 1' mechanisms. Complexity of the traditional CA problems forced a focus on individual, affective and rational cognition and little effort has been put into the development of CAA in social simulations. Paradoxically, this focus on individual not only limited CAA capabilities in group social simulation but also in modeling of individuals. 'System 1' manifests not only as affective dispositions manifesting in separation, but also in normative and cultural patterns like interpretations of social contexts, taboos or subscriptions of group identities and related stereotypes. Those elements are an important factor even in individual interaction, but are easier to omit in individual agents than other features like affect or personality. 'Studies with synthetic characters have so far infrequently considered the link between behavior and culture, it may become an invisible background; directly encoded into the design' [5].

3.4 Normative Multi-agent Systems in Computational Sociology

Despite high level of sophistication CA and CAA are rarely used to address socio-cognitive problems or to model related 'System 1' non-rationality. In this section we investigate how those issues are approached in a domain directly addressing human-like social phenomena interaction of multiple entities i.e. Multi-agent Systems (MAS) applied to Computational Sociology.

3.4.1 Paradigm issues in Social Sciences

Traditional Social Sciences remain torn by the fact that culture is central to them and yet it defies rigorous and fruitful research [11]. As a result, many branches of social science surprisingly downplay 'culture' as a term (implicitly or explicitly), while easy quantification of economical or structural factors leads to easily verifiable hypothesis, concept of culture became a 'conceptual garbage can' for poorly defined mechanisms [45]. Its use to explain any effects by trivially labeling them 'a cultural influence' has made the term unpopular, as explaining nothing in reality. It has been argued that this inability to properly address culture is a key factor keeping branches of social science 'fractionated , each speaking their own, often mutually unintelligible languages and holding assumptions and theories that are mutually incompatible' [45].

Ontologically, socio-cultural properties emerge from psychological ones. Inability of social sciences to capture this with sufficient formality results in disparity between description levels. This leads to macro-scale models representing culture 'as an object' with underrepresented emergent, process-like aspects, leaving out accumulated micro-scale psychological knowledge. While the fear of reductionist issues (i.e. reluctance to reduce sociology to psychology) plays a role, the micro-macro gap must be primarily attributed to limitations of the traditional verbal models and technical difficulties of creating quantitative, computational ones, surpassing expertise of social scientists.

3.4.2 Promise of computational social sciences. Normative Multi-agent Systems

At least since Axelrod [15] computational models are seen as a potential platform to unify and formalize the description of culture. Computationally, perhaps the most promising methodologies are based on a concept of agenthood, particularly normative Multi-agent Systems (MAS) has been explored greatly. As Neuman puts it 'cognitive architecture of the agents provides an objective reconstruction of the subjective dimension of norms. This provides a bridge between positivistic and phenomenological theory traditions. On the other hand, the inclusion of norms into the framework of the principally individualistic approach of agent based modeling provides an attempt to overcome the classical micro-macro divide.' [49].

In MAS, norms are on the one hand seen as pragmatic tools of distributed control assuring properties of a designed system and on the other as providing a bottom-up, individualistic explanation of social phenomena, when used in Computational Sociology. The term 'norm' is ambiguous in MAS and its intra-agent representation is not standardized [49]. Most popular choice are *constraints* i.e. hard limitations on agent actions designed off-line or *obligations* i.e. explicit conditional prescriptions incorporated e.g. in BOID architectures [7]. Rarely more complex mental objects are mentioned, usually *moral norms* and *values* [8] that accumulate all relevant cognitive aspects ill-represented by 'it-then' rules. Even when a norm is not simply seen as an external behavior regularity (e.g. 'all agents use right side of the road') the cognitive complexity of its representation is low compared to the mechanisms postulated in sociology.

3.4.3 Normative Multi-agent Systems and human-like non-rationality

There is a tendency in MAS, going back to Epstein [15] to identify culture with a) *norm emergence* as a behavior pattern propagation, b) *external products of a)* like trade and relation networks, production and consumption patterns, life span or housing distributions [15]. An implicit assumption that this reasonably approximates culture is based on a broad application of norms in psychology, as 'cultural product including values, customs, traditions etc.' [58]. This is however unjustified, considering narrow redefinition of the term in MAS. The fact that abstract conditional rules are not a good model of the innate habitual, affective and other complex particularities of normative mechanisms discussed in social psychology is usually silently ignored.

Externalistic focus and basic assumption of MAS to create complexity with interaction of simple agents promote cognitive over-simplicity. It also has trivial technical justification: agents with complex cognitive components are hard to model, implement, parametrize and validate. Constraints and obligations are easy to use, but 'it requires a lot to transform morality into a computational approach. It is still an open research question how this can be realized' [49]. This is even more true in a broader case of culture. Attempts to address this issue exist outside the MAS mainstream but are relatively limited. Breen [8] uses 'metanorms' to explicitly separate culture, and norms and to show how 'culture affects the possibility of normative changes, in particular the acceptance of policies' [8]. Kochanowicz [33, 34] proposes a model for MAS uniting norms, culture and personality on a level of social context cognition. Dignum et. al. propose increase in social context awareness, introduction of identity management and mention relevance of Dual process theory [17]. Despite similar attempts, psychological mechanisms of culture emergence and internalization are both rare and highly non-trivial - the solution to the posted problem exceeds traditional MAS expertise.

4. POSSIBLE RESEARCH DIRECTIONS

Based on the state of the major related fields with relation to human-like non-rationality and after discussing their limitations and potential in previous section, this section presents ideas and possible directions of future development.

4.1 CAA-MAS unification. Closing immergence-emergence loop

The starting observation is, that there is a great need for theoretical and implementational unification of the externalist, objective approach to normative-cultural phenomena represented by MAS or organization modeling with the internalist, cognitive present in CAA. MAS directly tackle the problem of socio-normative emergence, but depth of this pursuit is limited by the psychological complexity of the used agents. On the other hand CAA provide complex, affectively and cognitively believable agents, but are designed and used for investigation of the individual not social and cultural dimensions. Strengths and limitations of MAS and CAA are then complementary and despite the fact that they are separate disciplines with their own goals, their unification would be beneficial for simulations of human-like non-rationality and could improve modeling of both group and individual behavior. Additionally, CAA and MAS share a challenge: both are yet to explicitly incorporate 'System 2' mechanisms of human-like non-rationality especially in social cognition.

On a theoretical level, postulated unification can be framed as a circular dependency between the components of emergence-immergence duality [3]: collective feature formation of the group from individual behavior features and cultural influence of the group on the individual cognition. As Conte et al. put it: 'The interplay between the mental and the social dynamics allows norms to emerge and change. Observable conformity is only the tip of the normative iceberg. The crucial dynamics lies in the minds of the agents, beneath the line of observation. Norms cannot emerge in society unless they previously immerge in the mind, i.e. get converted into mental representations. Agents abiding with norms, or violating them, act on a set of specific, norm-related, mental representations' [14]. In both MAS and CAA, emergence-immergence loop is then 'broken', but the reasons are opposite. CAA provide complexity of cognitive models that could support culture immergence, but as they are not designed for and used in multi-agent, culture forming interactions this does not take place. MAS focus directly on the emergence but simple agents cause insufficient immergence: despite great popularity of norm simulations, cultural-normative immergence in cognition is still marginal. As Neumann puts it, 'future work could profit from a finer-grained resolution of internal processes of normative reasoning based on explicit representations of norms. While agent-based modeling has reached a substantial understanding of inter-agent processes, an investigation of the recursive impact of inter- and intra-agent processes is still in its fledgling stages' [48].

4.2 Formalizing human-like non-rationality

On a practical level, essential question is: what form should 'System 1' immergence take, specifically? It is a great research question, nevertheless some initial directions and ideas are proposed in this section. Most generally put, to model human-like non-rationality one must understand it in terms of the cognitive evolution. Human intelligence is not a general problem solver, but a historical and specific response to a particular survival task. Specifically, 'System 1' cultural immergence mechanisms were developed under evolutionary pressure to enable socialization in groups, shaping features of cognition to the degree that made many evolutionary anthropologists see this process, as one of the main reasons for the astonishing human brain development [19]. This implies at least three types of evolutionary contributions to non-rationality that models need to account for.

First, as evolution is based on the selection of genes not individuals, it promotes those non-rational mechanisms in individuals that serve the survival of the population sharing given genes, even at the cost of that individual (making those mechanisms irrational from the egoistic perspective). Those mechanisms include *explicitly altruistic or pro-social needs and capabilities*, but also more subtle *mechanisms of group synchronization*, improving its coordination and functioning and thus survival. Those mechanisms include needs and abilities to perceive, process and produce social structure, hierarchy or rules. Those mechanisms probably also provided a cognitive base for what later became morality, informal laws, religion or fashions. Those *group-centered 'System 1' needs and abilities of their realization* are a primary source of non-rational phenomena that will not emerge from an interaction of rational agents with egoistic goals.

Second implication is that some cognitive mechanisms effective in the past are suboptimal in modern contexts. Their particularity must also be included in the relevant models. This includes not only the clearly damaging tendencies like extreme anger, urges to kill or rape, that were valid means of protecting or spreading genes in the past and have opposite effect currently, but also less dramatic consequences of our genetic heritage. It could even be argued that entire human affective constitution is an atavistic answer to problems better addressed with reason, although such statement would probably be too strong.

Finally, aspect is related to the fact that evolution is a random process constrained by its 'medium' i.e. organic matter with physical limitations in terms of speed of operation, adaptivity etc. Notably 'inertia' and habitual nature of human decision making is less frequently mentioned then its resource (temporal and computational) limitations.

4.2.1 Internal reconstruction of social concepts

First specific implication of the above points is that human-like non-rationality will not only require explicit norm representations (like Neumann and others suggested), but a full redefinition of social concepts like 'role', 'group' etc. from organizational to a cognitive dimension. This is needed, as they are a base of the mentioned 'System 1' pro-social and group synchronizing capabilities.

For example, while in MAS 'role' usually represents objective, external agent differentiation within an organization, it must also be present as a subjective, cognitive dimension, based on internal image of a role as a psychological, normative 'archetype' that may fit the objective functional or organizational role, or not. Individual, internal role definitions may vary and change substantially, even within one group, while formal and external organizational aspects emerge from them and via synchronization-penalization mechanism that ensure that subjective variations do not endanger the group. Analogously norms, group identities etc. should be reintroduced from objective to subjective space.

4.2.2 'System 1' social context cognition

Above internal representations are the building blocks of human-like abilities realizing evolutionary promoted group synchronization, notably human-like 'System 1' ability to constantly maintain appropriate social context understanding. Humans flexibly adopt behavioral and normative standards to the circumstances of the social situation in a subtle, intuitive and subconscious manner and even individuals declaring a monolithic formal moral code, are influenced by this process. While general formal values or norms are often valid in most contexts, informal normative dispositions are characterized by context dependence, inconsistencies and fluid change, that implies a non-trivial management system based on a social context monitoring, rather than a static, uniform norm hierarchy, as it is the case in current CAA or MAS models.

The details of such social context cognition model are an open question. One possible source of the relevant components could for example be a Dramaturgical Theory (**DT**) [9]. According to its original, sociological formulation, all human behavior is inherently social and identity is not a stable, independent entity, but is constantly altered according to the specific contexts. More precisely, innate human social cognition is best described using terminology of a stage performance. The concept of a 'role' is central in DT, as a norm aggregating entity, not as an external, institutional group function atom, but an individually constructed, normative unit, constantly defined anew. Both role definitions and their subscriptions to agents are subjective, possibly conflicting with the views of other agents and may change. Similarly 'scene' with 'social scripts' refers to informal classes of social situations, with norms, goals, rituals, plans or emotional settings, subjectively perceived as acceptable in a given context. Individual content is constantly synchronized via social feedback, making it inter-subjective rather than purely subjective. Some terms relevant to the 'dramaturgical context cognition' can be found in some of the existing architectures (especially 'social role'), but are only rarely grouped to form a complete 'System 1' social context cognition system usable in CAA, see [35, 33, 34]. Non-exhaustive list of concepts to consider in such internal context cognition model includes: *group stereotypes* representing collective characteristics, the nature of group's relationships and atmosphere e.g. 'friends', 'family', 'religious' or 'militaristic' group; *role archetypes*, standing for the informal agent function; *social situations* representing intuitive social scenes; *values*, abstract qualities like 'virtue', 'honor'; *internal norms* pointing to socially acceptable behavior and socially imposed behavior limitations.

4.2.3 Implicit culture. From symbolic models to partly connectionist hybrids

Formal and explicitly formulated concepts and rules are only a narrow class of both intra and inter-agent normative, social and cultural signals. Externally, culture manifests both formally as laws-rules-norms and as informal patterns of behavior or feedback. Also internally, it immerges both as formal concepts and as informal, decentralized features of emotionality, temperament or implicit social expectations, often subconsciously obtained and used. Therefore, predominantly used formal predicate based models may not be optimal in handling 'System 1' aspects of those signals. Instead, connectionist representations may be better suited for modeling sub-symbolic, fuzzy, decentralized and associative nature of informal culture and thus social agents may benefit from incorporating both approaches.

5. SUMMARY AND CONCLUSIONS

This paper presented a survey of work in computational disciplines related to human-like non-rationality modeling, including: Social Signal Processing, Cognitive Architectures, Affective Computing, Human-Like Agents and Normative Multi-agent Systems with analysis of their current limitations and potential in contributing to the ability to adequately address human-like non-rationality. To enable precise presentation, key terms were clarified and background of the rationalistic perspective on the behavior simulation was presented. Finally, a proposal of directions and ideas for future development was discussed.

6. ACKNOWLEDGMENTS

This research, which is carried out at BeingThere Centre, collaboration among IMI of Nanyang Technological University (NTU) Singapore, ETH Zurich, and UNC Chapel Hill, is supported by the Singapore National Research Foundation (NRF) under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office (IDMPO).

7. REFERENCES

[1] C. Adam, A. Herzig, and D. Longin. A logical formalization of the occ theory of emotions. *Synthese*, 168(2):201–248, 2009.

[2] J. R. Anderson. *How can the human mind occur in the physical universe?* Oxford University Press, 2007.

[3] G. Andrighetto, M. CampennÃň, F. Cecconi, and R. Conte. *The complex loop of norm emergence: A simulation model*, pages 19–35. Springer, 2010.

[4] M. B. Arnold. Emotion and personality. 1960.

[5] R. Aylett, N. Vannini, E. Andre, A. Paiva, S. Enz, and L. Hall. But that was in another country: agents and intercultural empathy. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 329–336, 2009.

[6] M. Bateson, D. Nettle, and G. Roberts. Cues of being watched enhance cooperation in a real-world setting. *Biology letters*, 2(3):412–414, 2006.

[7] G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Proceedings of the*

second international joint conference on Autonomous agents and multiagent systems. ACM, 2003.

[8] J. M. Breen, G. D. Tosto, F. Dignum, and G. J. Hofstede. Linking norms and culture. In *Culture and Computing (Culture Computing), 2011 Second International Conference on*, pages 9–14. IEEE, 2011.

[9] D. Brissett and C. Edgley. *Life as theater: a dramaturgical sourcebook*. Transaction Publishers, 2005.

[10] J. Broekens, D. Degroot, and W. A. Kosters. Formal models of appraisal: Theory, specification, and computational model. *Cognitive Systems Research*, 9(3):173–197, 2008.

[11] M. J. Brown. *Explaining culture scientifically*. University of Washington Press, 2008.

[12] C. Castelfranchi and E. Lorini. Cognitive anatomy and functions of expectations. Citeseer.

[13] H.-Q. Chong, A.-H. Tan, and G.-W. Ng. Integrated cognitive architectures: a survey. *Artificial Intelligence Review*, 28(2):103–130, 2007.

[14] R. Conte, G. Andrighetto, and M. CampennÃň. *The immergence of norms in agent worlds*, pages 1–14. Springer, 2009.

[15] J. S. Dean, G. J. Gumerman, J. M. Epstein, R. L. Axtell, A. C. Swedlund, M. T. Parker, and S. McCarroll. Understanding anasazi culture change through agent-based modeling. *Dynamics in human and primate societies*, pages 179–205, 2000.

[16] O. E. Dictionary. Oxford english dictionary online. *Mount Royal College Lib., Calgary*, 14, 2004.

[17] F. Dignum, V. Dignum, C. Jonker, and R. Prada. Situational deliberation; getting to social intelligence. *Computational Social Science and Social Computer Science: Two Sides of the Same Coin*, 2014.

[18] W. Duch, R. J. Oentaryo, and M. Pasquier. Cognitive architectures: Where do we go from here? *Frontiers in artificial intelligence and applications*, 171:122, 2008.

[19] R. I. Dunbar. The social brain hypothesis. *brain*, 9:10, 1998.

[20] S. Epstein. The stability of behavior: I. on predicting most of the people much of the time. *Journal of personality and social psychology*, 37(7):1097, 1979.

[21] H. J. Eysenck. Dimensions of personality: 16, 5 or 3?âĂŤcriteria for a taxonomic paradigm. *Personality and individual differences*, 12(8):773–790, 1991.

[22] N. H. Frijda, P. Kuipers, and E. Ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of personality and social psychology*, 57(2):212, 1989.

[23] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[24] H. Gunes. Automatic, dimensional and continuous emotion recognition. 2010.

[25] M. G. Haselton, D. Nettle, and D. R. Murray. The evolution of cognitive bias. *The handbook of evolutionary psychology*, 2005.

[26] M. John Stuart. On the definition of political economy, and on the method of investigation proper to it. *Essays on some unsettled questions of Political Economy*, page 326, 1844.

[27] A. Kagian, G. Dror, T. Leyvand, I. Meilijson, D. Cohen-Or, and E. Ruppin. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision research*, 48(2):235–243, 2008.

[28] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[29] G. A. Kaminka. Curing robot autism: A challenge. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 801–804. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[30] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, 1(1):137–159, 2008.

[31] D. E. Kieras and D. E. Meyer. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human-computer interaction*, 12(4):391–438, 1997.

[32] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on*, 4(1):15–33, 2013.

[33] J. Kochanowicz, T. Ah-Hwee, and D. Thalmann. Enabling generative, emergent artificial culture. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[34] J. Kochanowicz, A.-H. Tan, and D. Thalmann. Beyond traits: Social context based personality model. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1529–1538. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[35] J. Kochanowicz, A.-H. Tan, and D. Thalmann. Dramaturgical and dissonance theories in explicit social context modeling for complex agents. *Computer Animation and Virtual Worlds*, 26(3-4):247–257, 2015.

[36] J. E. Laird. *The Soar cognitive architecture*. MIT Press, 2012.

[37] P. Langley and D. Choi. A unified cognitive architecture for physical agents. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1469. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[38] P. Langley, J. E. Laird, and S. Rogers. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160, 2009.

[39] J. Lin, M. Spraragen, and M. Zyda. Computational models of emotion and cognition. In *Advances in Cognitive Systems*. Citeseer, 2012.

[40] R. P. Marinier III and J. E. Laird. Toward a comprehensive computational model of emotions and feelings. In *ICCM*, pages 172–177.

[41] S. Marsella and J. Gratch. Ema: A computational model of appraisal dynamics. *Agent construction and emotions*, pages 601–606, 2006.

[42] S. Marsella, J. Gratch, and P. Petta. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, pages 21–46,

2010.
[43] D. P. McAdams. What do we know when we know a person? *Journal of personality*, 63(3):365–396, 1995.
[44] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.
[45] A. Mesoudi. *Cultural evolution: how Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press, 2011.
[46] J.-J. Meyer. Reasoning about emotional agents. *Artificial Intelligence Preprint Series*, 44, 2008.
[47] W. Mischel and Y. Shoda. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246, 1995.
[48] M. Neumann. Homo socionicus: a case study of simulation models of norms. *Journal of Artificial Societies and Social Simulation*, 11(4):6, 2008.
[49] M. Neumann. The cognitive legacy of norm simulation. *Artificial Intelligence and Law*, 20(4):339–357, 2012.
[50] G. W. Ng. *Brain-mind machinery: brain-inspired computing and mind opening*. World Scientific, 2009.
[51] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.
[52] M. Pantic, R. Cowie, F. DâĂŹErrico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli. Social signal processing: the research agenda. In *Visual analysis of humans*, pages 511–538. Springer, 2011.
[53] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri. Multimodal support to group dynamics. *Personal and Ubiquitous Computing*, 12(3):181–195, 2008.
[54] R. W. Picard. Affective computing. 1995.
[55] A. S. Rao, M. P. Georgeff, et al. Bdi agents: From theory to practice.
[56] R. Reisenzein, E. Hudlicka, M. Dastani, J. Gratch, K. Hindriks, E. Lorini, and J.-J. C. Meyer. Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange. *Affective Computing, IEEE Transactions on*, 4(3):246–266, 2013.
[57] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *Multimedia, IEEE Transactions on*, 11(7):1373–1380, 2009.
[58] M. Sherif. The psychology of social norms. 1936.
[59] D. J. Sollenberger and M. P. Singh. *Architecture for affective social games*, pages 79–94. Springer, 2009.
[60] R. Sun. The clarion cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and multi-agent interaction*, pages 79–99, 2006.
[61] R. Sun. *Cognitive architectures and multi-agent social simulation*, pages 7–21. Springer, 2009.
[62] E. L. Thorndike. A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29, 1920.
[63] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.
[64] Z. Yumak, J. Ren, N. M. Thalmann, and J. Yuan. Modelling multi-party interactions among virtual characters, robots, and humans. *Presence: Teleoperators and Virtual Environments*, 23(2):172–190, 2014.
[65] G. Zacharias, J. MacMillan, and S. B. Van Hemel. *Behavioral modeling and simulation: from individuals to societies*. National Academy Press, 2008.