COGNTIVE 2022

The Fourteenth International Conference on Advanced Cognitive Technologies and Applications

ISBN: 978-1-61208-950-8

April 24 - 28, 2022

Barcelona, Spain

COGNITIVE 2022 Editors

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

COGNITIVE 2022

Forward

The Fourteenth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2022), held on April 24 - 28, 2022, targeted advanced concepts, solutions and applications of artificial intelligence, knowledge processing, agents, as key-players, and autonomy as manifestation of self-organized entities and systems. The advances in applying ontology and semantics concepts, web-oriented agents, ambient intelligence, and coordination between autonomous entities led to different solutions on knowledge discovery, learning, and social solutions.

The conference had the following tracks:

- Brain information processing and informatics
- Artificial intelligence and cognition
- Agent-based adaptive systems
- Applications
- Autonomous systems and autonomy-oriented computing
- Hot topics on cognitive science

Similar to the previous edition, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the COGNITIVE 2022 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to COGNITIVE 2022. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the COGNITIVE 2022 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope COGNITIVE 2022 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of cognitive technologies and applications. We also hope that Barcelona provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

COGNITIVE 2022 General Chair

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

COGNITIVE 2022 Steering Committee

Charlotte Sennersten, CSIRO Mineral Resources, Australia
Jayfus Tucker Doswell, The Juxtopia Group, Inc., USA
Roberto Saracco, IEEE New Initiative Committee Chair, Italy
Thomas Ågotnes, University of Bergen, Norway
Muneo Kitajima, Nagaoka University of Technology (Emeritus), Japan

COGNITIVE 2022 Publicity Chair

Lorena Parra, Universitat Politecnica de Valencia, Spain
Javier Rocher, Universitat Politècnica de València, Spain

COGNITIVE 2022

Committee

COGNITIVE 2022 General Chair

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

COGNITIVE 2022 Steering Committee

Charlotte Sennersten, CSIRO Mineral Resources, Australia
Jayfus Tucker Doswell, The Juxtopia Group, Inc., USA
Roberto Saracco, IEEE New Initiative Committee Chair, Italy
Thomas Ågotnes, University of Bergen, Norway
Muneo Kitajima, Nagaoka University of Technology (Emeritus), Japan

COGNITIVE 2022 Publicity Chair

Lorena Parra, Universitat Politecnica de Valencia, Spain
Javier Rocher, Universitat Politècnica de València, Spainin

COGNITIVE 2022 Technical Program Committee

Witold Abramowicz, University of Economics and Business, Poland
Thomas Agotnes, University of Bergen, Norway
Vered Aharonson, University of the Witwatersrand, Johannesburg, South Africa
Sadam Al-Azani, King Fahd University of Petroleum & Minerals, Saudi Arabia
Luis Alfredo Moctezuma, Norwegian University of Science and Technology, Trondheim, Norway
Piotr Artiemjew, University of Warmia and Masuria in Olsztyn, Poland
Divya B, SSNCE, India
Petr Berka, University of Economics, Prague, Czech Republic
Ateet Bhalla, Independent Consultant, India
Mahdi Bidar, University of Regina, Canada
Guy Andre Boy, CentraleSupélec, LGI, Paris Saclay University / ESTIA Institute of Technology, France
Dilyana Budakova, Technical University of Sofia - Branch Plovdiv, Bulgaria
Valerie Camps, Paul Sabatier University - IRIT, Toulouse, France
Yaser Chaaban, Leibniz University of Hanover, Germany
Olga Chernavskaya, P. N. Lebeved Physical Institute, Moscow, Russia
Helder Coelho, Universidade de Lisboa, Portugal
Igor Val Danilov, Academic Center for Coherent Intelligence, Latvia
Angel P. del Pobil, Jaume I University, Spain
Soumyabrata Dev, University College Dublin, Ireland
Jerome Dinet, University of Lorraine, France
Piero Dominici, University of Perugia, Italy

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Language and Image in Behavioral Ecology

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

Jérôme Dinet
Université de Lorraine, CNRS, INRIA, Loria
Nancy, France
Email: jerome.dinet@univ-lorraine.fr

Clelie Amiot
Univ. de Lorraine, CNRS, INRIA, Loria
Nancy, France
Email: clelie.amiot@univ-lorraine.fr

Capucine Bauchet
Univ. de Lorraine, DANE Nancy-Metz
Nancy, France
Email: capucine.bauchet@univ-lorraine.fr

Hanna Verdel
Univ. de Lorraine, DANE Nancy-Metz
Nancy, France
Email: hanna.verdel@univ-lorraine.fr

***Abstract*—Ideas are created in one's mind through cognitive processes after obtaining perceptual stimuli either by hearing or reading words or by seeing images. They should have different representations depending on their origin of information, i.e., words or images, and the cognitive processes for dealing with them. The comparison between these processes is often labeled by the terms, "word and wordless thought" and there is a strong argument that favors wordless thought. The purpose of this paper is to compare the two cognitive processes for words and images by applying the state of the art cognitive architecture, the Model Human Processor with Realtime Constraints (MHP/RT) proposed by Kitajima and Toyota, developed for understanding behavioral ecology of human beings. This study shows that the perceived dimensionality of images is larger than that of words, which leads to the conclusion that the number of discriminable states for images is an order of magnitude larger than that of words, and due to this, image-based processing can store information about absolute times in memory but word-based processing cannot. This should lend significantly larger expressive power to image-based processing. It is argued that the loss of reality in word-based processing results in significant implications for the development of globalization and the illusion of mutual understanding in word-level communications.**

Keywords–Word and wordless thought; Cognitive architecture; MHP/RT; Loss of reality.

I. INTRODUCTION

There is a teaching that says, "No matter how many times you *listen* to it, you can't actually *see* it even once. You should see everything with your own eyes." This is what Zhao Chongguo, a general of the Former Han, said as he introduced the following passage from a volume entitled "A History of Zhao Chongguo" in the Book of Han or History of Former Han:

> The Han Emperor asked Zhao Chongguo about the strategies and forces needed to quell the rebellious Tibetan nomads. Zhao Chongguo asked for forgiveness, saying, "Since it is difficult to formulate a strategy in a distant place, I would like to go to the site and draw a map of what I actually saw and tell a trick."

This somewhat abstract teaching is also demonstrated in the well-known sentence, "it is better to see it with your own eyes than to hear it a hundred times," which is an expression that can be translated into actions in a more understandable way. In this expression, hearing something a hundred times is equated with seeing it once. The following two translations are also derived from this: "a picture is worth ten thousand words" and "seeing is believing."

Regarding the first translation, Larkin and Simon [2] posed the following research question from the cognitive scientific viewpoint: "Why a diagram is (sometimes) worth ten thousand words." In order to find the answer, they assumed a situation where the same amount of information was expressed by diagrams and sentential paper-and-pencil representations and examined their characteristics from the viewpoint of human information processing to see if the amount of knowledge that one can acquire by *seeing a diagram once* is equivalent to the amount of knowledge that one can acquire by *hearing ten thousand words* that explain that diagram. Words and wordless thought are controversial issues in philosophy. One position identifies words or language with cognition, stating that an idea cannot be conceived other than through the word and only exists by the word, while wordless thought corresponds to the cognitive processes that are invoked when seeing a diagram. Jacques Salomon Hadamard, a distinguished French mathematician, described his own mathematical thinking as largely *wordless*, often accompanied by mental images that represent the entire solution to a problem. In his book entitled, "Psychology of Invention in the Mathematical Field [3]," he tried to report and interpret observations, either personal or gathered from other scholars engaged in the work of invention.

Regarding the second translation, "seeing is believing" implies that the state of believing something, which can be rephrased as the behavior of accepting the truth, reality, or validity of something (e.g., a phenomenon or a person's veracity), could be reached by seeing it. The perceptual and cognitive processes that occur between seeing the site of the rebellion and accepting the reality of the rebellion in the case of Zhao Chongguo, can be mapped on Two Minds [1][4], which suggests that human behavior emerges as the result of competition between the dual processes of System 2, which is a slow conscious process for deliberate reasoning with feedback control, and System 1, which is a fast unconscious process for intuitive reaction with feedforward control for connecting perception and motor. The section surrounded by the dotted line rectangle in Figure 1 is a modified version

 ISBN: 978-1-61208-950-8

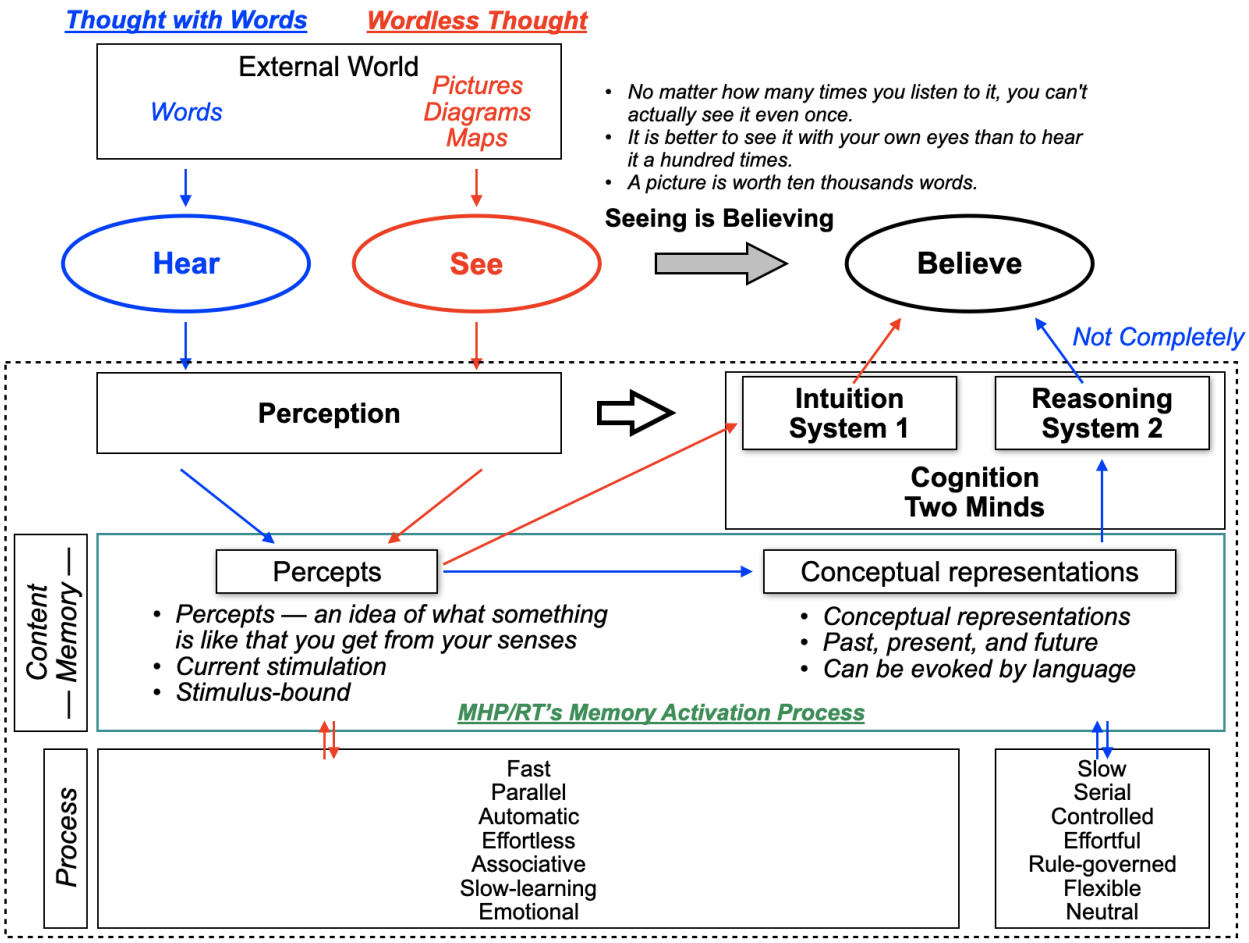


Figure 1. Word and wordless thought processes represented on the modified version of the figure used by Kahneman to explain Two Minds [1, Figure 2].

of the Figure used by Kahneman to explain Two Minds [1, Figure 2]. In Figure 1, the process that should correspond to "Seeing is Believing" is characterized as "Wordless Thought." It starts by seeing pictures, diagrams, or maps provided in the external world, followed by forming a percept which is an idea of what something is like that you get from your senses and from intuitively reaching the state of believing through System 1. The behaviors expressed in the teachings referred to above, i.e., "hearing ten thousand words," "to hear something a hundred times," or "to listen to something many times," are shown as "Thought with Words" in Figure 1. This process starts by hearing words, followed by forming a percept and reaching a state of not-completely-believing after extensive and deliberate reasoning processes through System 2.

In this paper, we aimed to dig deeper into the teaching in the Book of Han or History of Former Han and its variants by applying state-of-the-art cognitive architecture. In particular, the difference between Wordless Thought and Thought with Words will be clarified from the viewpoint of the difference in how layered structured memory is activated. The particular cognitive architecture we used for the analysis was the Model Human Processor with Realtime Constraints (MHP/RT) [5][6], which was applied to a variety of phenomena related to action selection and memory processes [7]–[15].

The remainder of this paper is organized as follows: Section II reviews MHP/RT focusing on the use of memories in the processing of System 1 and System 2, which is critical to understand the differences between image-based and word-based processing. Section III demonstrates that the distinction between word and wordless thought is one of the most ancient debates in psychology and has its roots in philosophical and educational considerations, and that several modern theories are based on this distinction. Section IV describes in detail how language and image are processed by MHP/RT focusing on how reality is guaranteed in these processes. Section V concludes the paper by summarizing the contents and pointing out the implications of the loss of reality in word-based communication in the development of globalization.

II. Model Human Processor with Realtime Constraints and Multi-dimensional Memory Frames

Kitajima and Toyota [6][16] constructed a comprehensive theory of action selection and memory, known as the Model Human Processor with Realtime Constraints (MHP/RT), that provides a basis for constructing any model for understanding human behavior (see Figure 2).

 ISBN: 978-1-61208-950-8

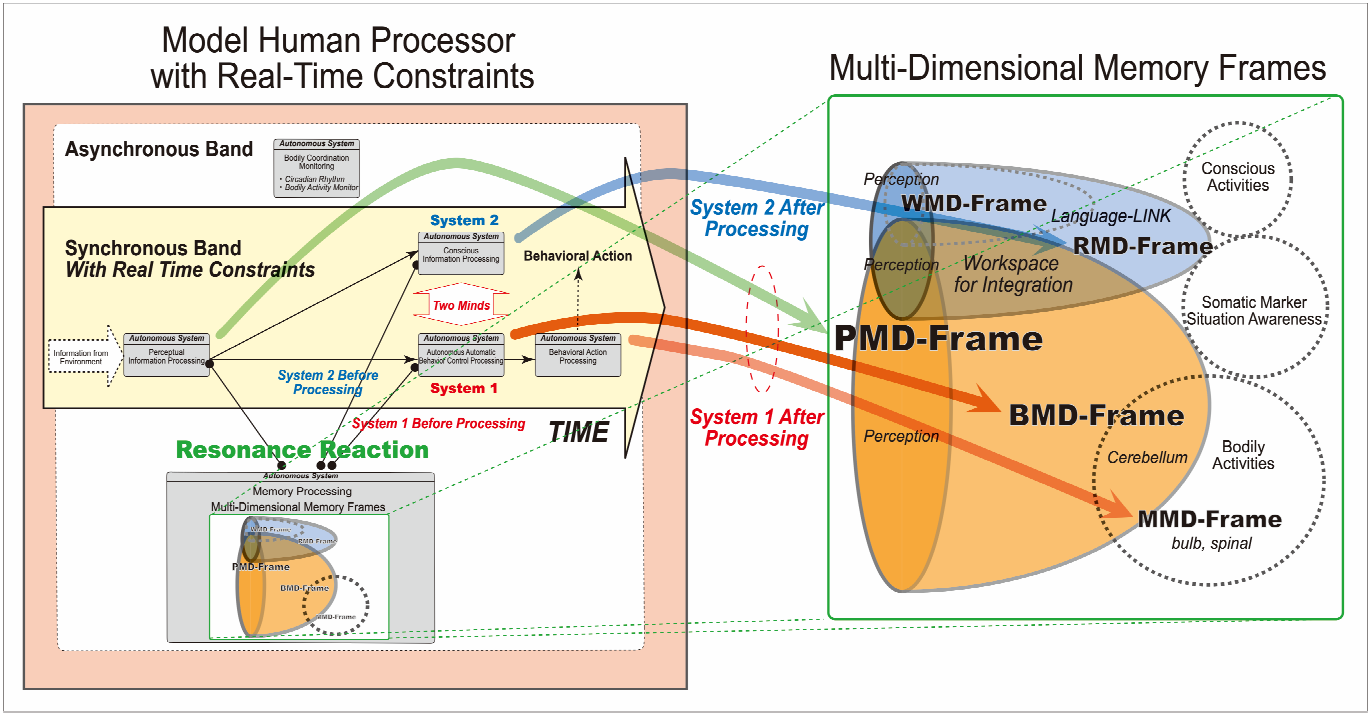


Figure 2. MHP/RT and the distributed memory system implemented as multi-dimensional memory frames (modified from [6, Figure 3]).

A. Outline of MHP/RT

The MHP/RT is an extension of the Model Human Processor proposed by Card, Moran, and Newell [17] that can simulate routine goal-directed behaviors. The process involved in action selection is a dynamic interaction that evolves in the irreversible time dimension. The purpose of MHP/RT is to explain the following three facts that underpin an understanding of the behavioral ecology of human beings:

1) The fundamental processing mechanism of the brain is Parallel Distributed Processing (PDP) [18], which is referred to as the Organic PDP (O-PDP) system in the development of MHP/RT.
2) Human behavior emerges as a result of competition between the dual processes of System 1, fast unconscious processes for intuitive reaction with feedforward control that connect perception with motor movements, and System 2, slow conscious processes for deliberate reasoning with feedback control. This is called Two Minds [1].
3) Human behavior is organized into 17 happiness goals [19].

B. Part 1: PCM Processes

MHP/RT consists of two parts. The first comprises cyclic PCM processes (Figure 2, left), in which PDP for these processes is implemented in hierarchically organized bands with characteristic operation times by associating relative times (not absolute) with the PCM processes that carry out a series of events that are synchronous with changes in the external environment. There is a gap between two adjacent bands; these two bands are non-linearly connected and therefore it is inappropriate to understand the phenomena that occur across these bands by constructing a linear model. The phenomena occur by connecting what happens in a band to what happens in its adjacent band non-linearly. A mechanism is required to connect the phenomena; MHP/RT suggests that this connection is provided by the resonance mechanism via the multi-dimensional memory frames.

C. Part 2: Multi-Dimensional Memory Frames

The bottom-left and right sections of Figure 2 show the autonomous memory system consisting of multi-dimensional memory frames of perception, motion, behavior, relation, and word. These memory frames store information associated with the corresponding autonomous processes defined in the PCM processes. The memory frames are subservient to the PCM processes because they do not exist unless the PCM processes exist.

The right section of Figure 2 shows the five memory frames and their relationship with the PCM processes. The following provides brief explanations of the respective memory frames.

- **WMD (Word MD)-frame** is the memory structure for language. It is constructed on a very simple one-dimensional array.
- **RMD (Relation MD)-frame** is the memory structure associated with the conscious information processing. It combines a set of BMD-frames into a manipulable unit.
- **BMD (Behavior MD)-frame** is the memory structure associated with the autonomous automatic behavior

control processing. It combines a set of MMD-frames into a manipulable unit.

- **PMD (Perceptual MD)-frame** constitutes perceptual memory as a relational matrix structure. It incrementally grows as it creates memory from the input information and matches it against the past memory in parallel.
- **MMD (Motion MD)-frame** constitutes behavioral memory as a matrix structure. It gathers a variety of perceptual information as well to connect muscles with nerves using spinals as a reflection point. In accordance with one's physical growth, it widens the range of activities the behavioral action processing can cover autonomously.

The memory frames have overlapping regions as follows: the PMD-frame overlaps with the WMD-, the RMD-, and the BMD-frames; the WMD-frame overlaps with the RMD-frame; and the BMD-frame overlaps with the MMD-frame. The PCM processes work for carrying out appropriate actions in response to the input stimuli. These actions are carried out when the corresponding portions of the MMD-frames are activated. The MMD-frames only overlap with the BMD-frame, and not with the RMD-frame or the WMD-frame, which System 2 operates with. This means that there is no direct path for System 2 to the MMD-frame to initiate any actions. System 2 can only indirectly contribute to the real actions via the BMD-frame which is connected to the MMD-frame.

D. Resonance as a Mechanism for Interaction Between PCM Processes and Memories

An important feature of the memory system is that it works *asynchronously* with the external environment. MHP/RT assumes that the *synchronous* PCM processes, including the perceptual system, System 1, System 2, the motor system, and the asynchronous memory system communicate with each other through a resonance mechanism. The concept of resonance has been borrowed from physics to describe the link between the asynchronous memory system and synchronous PCM processes. As Dinet et al. [12] suggested, apprehension of psychological phenomena using concepts borrowed from physics is useful because the majority of the interactions, including psychological interactions, between humans and the environment (social or physical environment) can be derived from physical processes.

Through the resonance process, the memory-frames work for the PCM processes to map perceptual information represented in the dimensionality of M to a motion represented in the dimensionality of N. In other words, the memory frames implement the $M \otimes N$ mapping from perception to motion via the resonance mechanism that connects the memories with the PCM processes. Figure 3 presents the relationships between the PCM processes shown at the bottom and the multi-dimensional memory frames shown at the top of the figure. It is important to note that System 1 has direct and parallel paths from perception to motion via the PMD-, the BMD-, and the MMD-frames, whereas System 2 does not in the $M \otimes N$ mapping. System 2 carries out serial processing along the paths from the PMD-frame to the WMD- and the RMD-frames, which are the memory for System 2. The results of System 2's processing could be transferred to the MMD-frame that enables actual actions through the overlaps between the RMD-frame and the BMD-frame. This can be described as follows: there are *direct* mappings of $M \otimes N$ for System 1 and there are *indirect* contributions of System 2's workings in these mappings, which is informally denoted as $M \otimes$(WORDS)$\otimes N$.

III. WORD AND WORDLESS THOUGHT

This section aims to demonstrate that (i) the distinction between word and wordless thought is one of the most ancient debates in psychology and has its roots in philosophical and educational considerations, and (ii) several modern theories (e.g., Dual Coding Theory and Cognitive Theory of Multimedia Learning) are based on this distinction.

A. The Philosophical Roots of the Debate

The distinction between word and wordless thought is believed to have inspired all cognitive and behavioral sciences that are interested in all forms of psychological explanations for the behavior of non-linguistic and linguistic creatures [20].

This distinction between word and wordless thought has been one of the bases of all educational manuals for several centuries. For instance, the educational pioneer, Jan Amos Komensky (alias Comenius), was the first author who, in the 17th century, proposed manuals that combined texts and pictures. His book entitled "Orbis Sensualium Pictus" (translated as "The world explained in pictures"), published in 1658 in Nuremberg, was the first widely used children's textbook with pictures. It was first published in Latin and German and later republished in many European languages, quickly spreading around Europe and becoming the definitive children's textbook for three centuries, with more than 200 editions published in twenty-six languages. This manual contains an extended summary of the world in 150 pictures with titles (Figure 4, left). All of the objects in the pictures are numbered and accompanied by parallel columns of labels and short sentences describing the numbered objects, categorized in different domains (zoology, religion, botany, etc.). More than 350 years later, as the right-hand section of Figure 4 shows, education manuals always have a similar appearance, where text and images are combined, because it is assumed that this combination has a positive impact on understanding and memory of information [21][22]. The main difference between manuals printed in the 17th century and modern manuals is that colors and typographical cues have been added due to progress in modern printing.

While some authors have been proposing materials that combine word and wordless thought for several centuries, psychological theories that explain the impact of this combination are more recent. For instance, Dual Coding Theory (DCT) has its roots in the practical use of imagery as a memory aid dating back 2,500 years [23][24]. Cognition, according to DCT, involves the activity of two distinct subsystems: (i) a verbal system specialized for dealing directly with language; and (ii) a non-verbal (imagery) system specialized for dealing with non-linguistic objects and events. The systems are assumed to be composed of internal representational units, called logogens and imagens, that are activated when one recognizes,

 ISBN: 978-1-61208-950-8

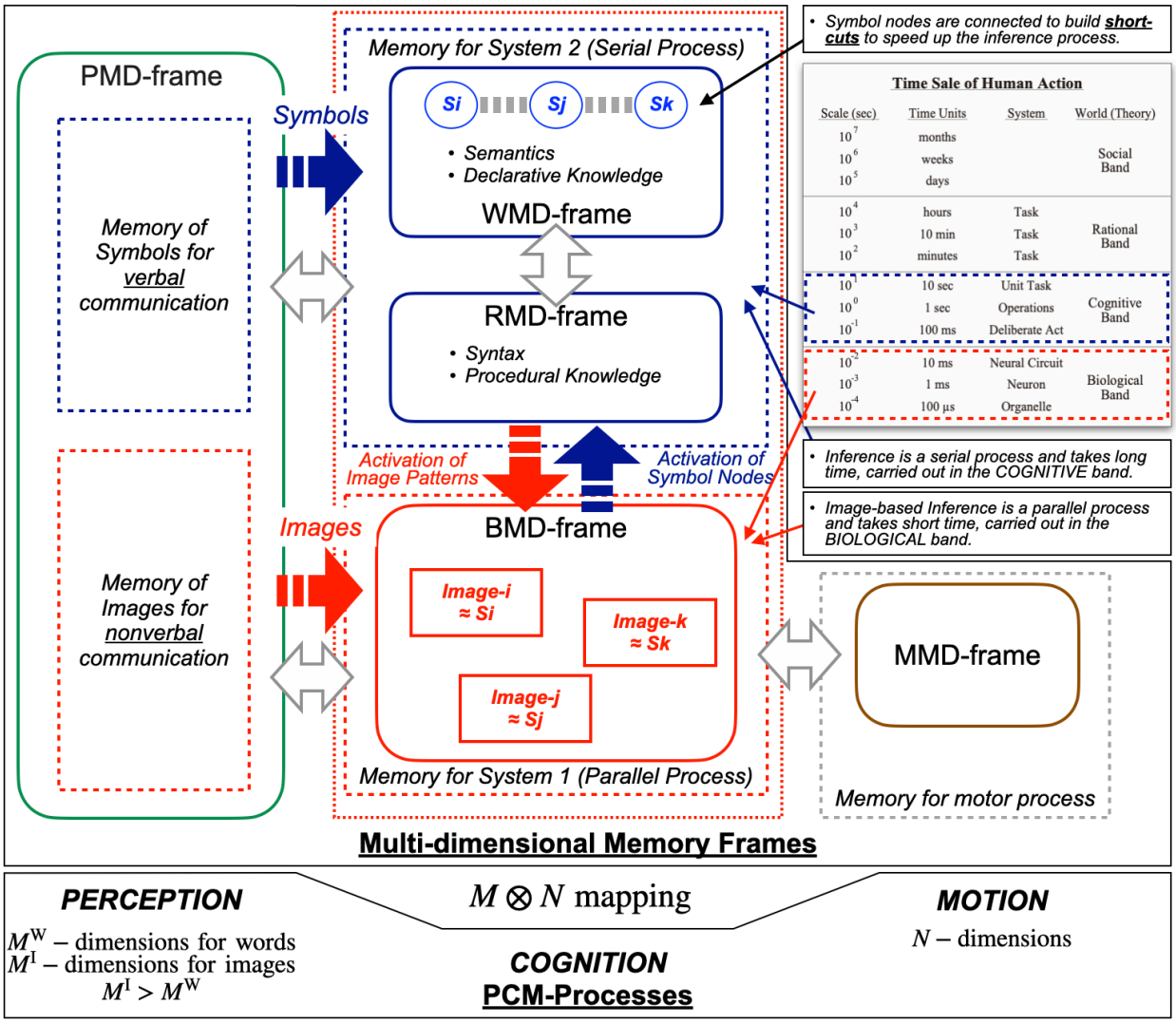


Figure 3. The relationships between the PCM processes shown in the bottom of Figure 3 and the multi-dimensional memory frames shown in the top of Figure 3.

manipulates, or just thinks about words or things. From a developmental point of view, dual coding development begins with the formation of a substrate of non-verbal representations and imagery derived from a child's observations and behaviors related to concrete objects and events, and relations among them. Language builds upon this foundation and remains functionally connected to it as referential connections are being formed, so that the child responds to object names in the presence or absence of the objects and begins to name and describe them (even in their absence). The events, relations, and behaviors are dynamically organized (repeated with variations) and thereby display natural syntax that is incorporated into the imagery as well. The natural syntax is enriched by motor components derived from the child's actions, which have their own patterns.

B. Modern Psychological Theories

A series of behavioral and neuropsychological studies provide further relevant support for this dual-channel approach (word versus wordless thought). For example, Thompson and Paivio [25] showed that object pictures and sounds had additive effects on memory, thereby supporting the DCT assumption that sensory components of multimodal objects are functionally independent. Similar effects have been demonstrated for combinations of other modalities. Brain scan studies have shown that different brain areas are activated by concrete and abstract words, as well as by pictures, as compared to words in comprehension and memory tasks (summarized in [23]). Other meta-analyses examined the most common loci of activation in fMRI and PET studies comparing abstract and concrete conceptual representations and support the dual-channel approach [26][27].

More recently, other theories related to psychology and education sciences have also been based on the distinction between word and wordless thought. For instance, the Cognitive Theory of Multimedia Learning, which is based on the Cognitive Load Theory and DCT, was developed after considering the previous theories, and is defined as the learning that

 ISBN: 978-1-61208-950-8

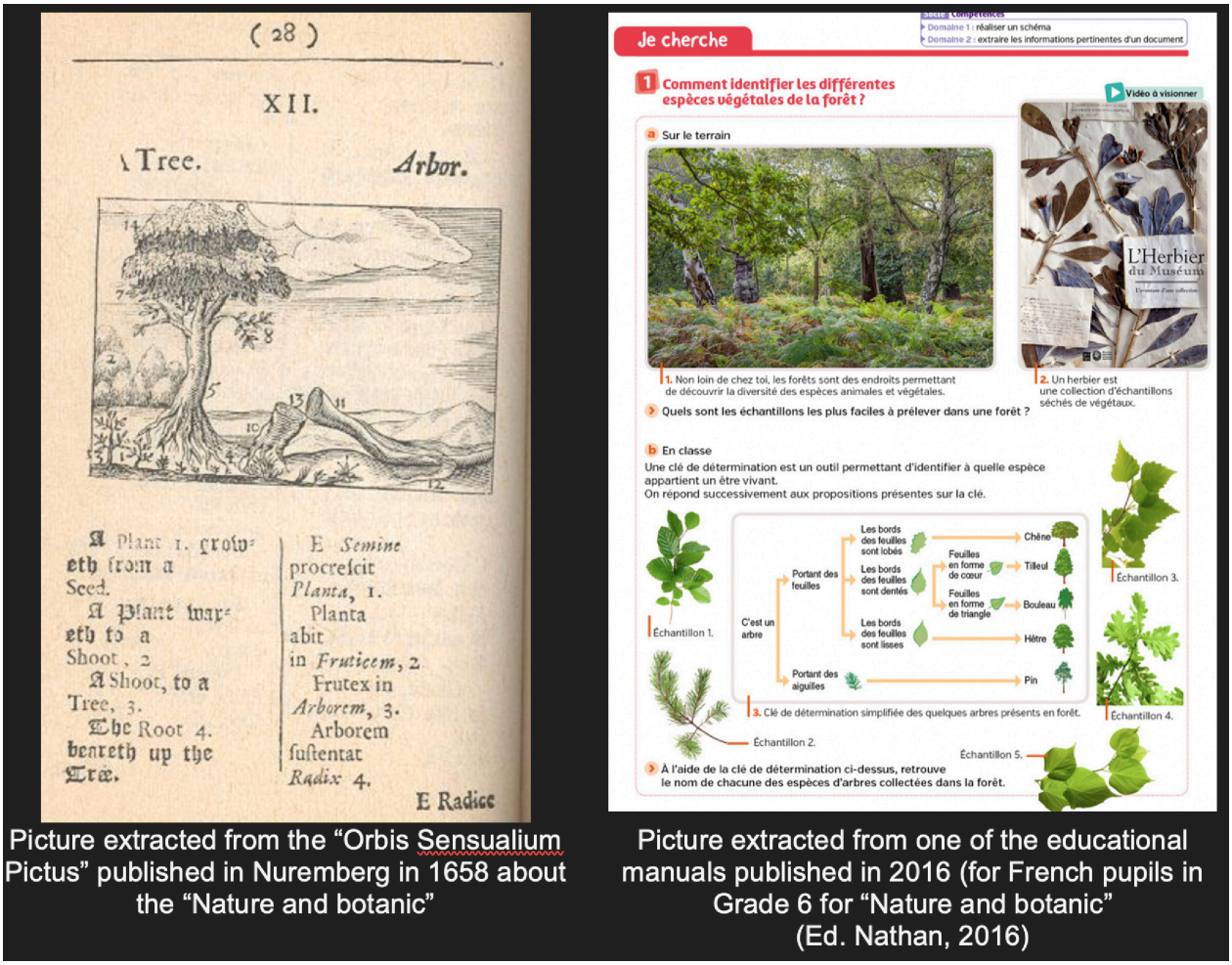


Figure 4. Educational manuals through the time, from the 17th Century to 21st Century.

is realized when constructing mental representations through pictures and words [21][28]–[30]. The Cognitive Theory of Multimedia Learning addresses how individuals process information and how they learn through multimedia approaches. It encompasses three fundamental assumptions:

1) People have two separate channels for processing visual and audio information.
2) Each channel has a limited amount of information per unit of time.
3) People experience active learning by accessing related information, organizing the selected information through mental structures, and integrating them with previous mental structures.

In other words, the dual-channel assumption is incorporated into the Cognitive Theory of Multimedia Learning by proposing that the human information-processing system contains an auditory/verbal channel and a visual/pictorial channel. According to Mayer [21][28]–[30], the relationship between the two channels is under the control of the user in the sense that users may be able to convert the representation for processing in the other channel if their cognitive capacity is sufficient. The limited capacity assumption (i.e., humans are limited as to the amount of information that can be processed in each channel at one time) is also very important for the Cognitive Theory of Multimedia Learning. Metacognitive strategies are techniques for allocating and adjusting the limited cognitive resources of the center executive (i.e., the system that controls the allocation of cognitive resources) and play a central role in all modern theories of intelligence [31].

C. Debate is Always Actual

Finally, we note that the debates between word and wordless language are at the core of recent studies investigating production and comprehension of a new generation of emoticons that have continued to grow in popularity and usage in both mobile communications and social media [32]. More precisely, because emojis have a distinctively social nature and are arguably extra-linguistic in origin, in that they are not part of the standard lexicon of any natural language, they have a special place in digital communication [33]. Today, emojis are pictographs that have become common units of expression. They are non-verbal cues for emotion (anger, joy, celebration) and similarly evocative expressions in digital communication, including text messages sent through smartphones and on social media platforms such as Twitter [32].

 ISBN: 978-1-61208-950-8

IV. Language and Image Processing in Behavioral Ecology

As an example of a situation where the teaching "seeing is believing" can be applied, we can use the sight of cherry blossoms falling. When a person actually sees this scene, s/he perceives it as a visual image. When s/he is provided with the description of the scene, e.g., "the beauty of cherry blossoms fascinates people, not only because of the blossoms themselves, but also because they are short lived" via printed or handwritten text on paper, or as an audible voice from a speaker or a person, s/he perceives it verbally as visual or auditory language. The perceived information is stored in working memory as cognitive frames for further processing [5, Figure 5]. Three things could follow, as indicated by the three lines leaving the box labeled "Perceptual Information Processing" in Figure 2:

1) The perceived information resonates with the multi-dimensional memory frames.
2) It is transferred to System 2 to initiate deliberate and conscious processing, e.g., inferences.
3) It is transferred to System 1 to initiate automatic and unconscious processing such as immediate emotional reactions for the voices.

While System 2 and System 1 are processing information, they can incorporate the activated portion of the memories through resonance. These processes are shown graphically by •——• in Figure 2 by connecting System 2 and System 1 with multi-dimensional memory frames. This section focuses on how MHP/RT uses memory to process languages and images in order to clarify the differences between these processes.

A. Processing Languages

This section focuses on how verbal inputs are processed by the PCM processes and the multi-dimensional memory frames.

1) Transferring Perceived Words to the WMD-frame: The perceived information presented verbally could resonate with the contents in the PMD-frame, i.e., the memory of symbols for verbal communication as shown at the top of the PMD-frame in Figure 3. The resonance would spread within the PMD-frame and transfer to the WMD-frame, where semantics of concepts is stored as declarative knowledge. The WMD-frame overlaps with the RMD-frame. These memory frames jointly advance language processing.

The following section examines in detail what would happen in the WMD-frame after the perceived information has arrived. Firstly, the structure of the WMD-frame, which is *functionally* constructed on a very simple one-dimensional array [34], should impose a strong restriction on the results of the resonance between the perceived information and the PMD-frame. Let L_i $(i = 1, \cdots)$ be a node in the WMD-frame and L_α be the node that receives the information from the PMD-frame. Since the memory related with L_α is constructed as a one-dimensional array, the activations originated from L_α could spread to $\{(L_{\alpha-1},\ L_{\alpha+1}),\ (L_{\alpha-2},\ L_{\alpha+2}),\ \cdots,\ (L_{\alpha-n},\ L_{\alpha+n})\}$ as time goes by along the connected nodes. After the PMD-frame establishes resonance with the perceived information, activation could spread in the WMD-frame with the overlapping node, L_α, as its origin.

2) Structure of the WMD-frame: The WMD-frame stores two types of language. The first type is "spoken language," i.e., *parole*, which defines spontaneously generated repetitive usage patterns of *phonetic symbols*. It is used as a means for exchanging information in the collective ecologies of humans. It should mirror human–human bodily interactions that are carried out under the structure of bodily functions [35]. The more frequently a sequence of phonetic symbols is used, the tighter it becomes to form a firm one-dimensional array of phonetic symbols. Possible interactions are restricted by the context in which they occur, which then restricts the range of spoken language that could emerge. This leads to construction of a thesaurus, which is a form of controlled vocabulary that seeks to dictate semantic manifestations of spoken language, and to make use of them while interacting with others. Semantic information could be hierarchically organized according to the levels of abstraction as information accumulates.

The second type is "written language," i.e., *langue*, which is a notational system with *logical symbols*, created by thoughtfully and evolutionarily developing spoken language for the purpose of efficient communication. The strong generality found in written languages leads to the establishment of grammar, i.e., *syntax*. The syntactic rules are stored in the RMD-frame, which could be activated by words stored in the WMD-frame using the overlap between them. This defines how System 2 carries out inference by utilizing the WMD-frame for semantics and the RMD-frame for syntax, which is a serial process and takes a long time, carried out in the cognitive band of Newell's time scale of human action [36, Fig. 3-3]. The one-dimensionality of the WMD-frame is the direct reflection of the nature of System 2's processing, which is serial processing. This means that only one node could be focused while performing inference. Therefore, the trajectory of inference processes could be represented as a series of one-by-one focused-on nodes.

B. Language vs. Image in Expressing Reality

1) Language Loses the Information Concerning Absolute Times: As described in Section IV-A, language realizes efficient human–human communication by generalizing repetitive usage patterns of phonetic or logical symbols. The process of generalization is called abstraction from the viewpoint of the degree of concreteness, or simplification from the viewpoint of the level of detail in expressing the situation. A generalized pattern represents wide variations of its instantiations in the real world. It can be used in human–human communication to express specific instantiations by using the generalized pattern that implies them.

As such, language is not appropriate for expressing reality, which is a collection of concrete instantiations in the real world. This is because language solely stores relationships, $\texttt{Rel}$, between events $E(T_1)$ and $E'(T_2)$ that happen at times T_1 and T_2, respectively. Let us suppose that $E(T_1)$ is an instance of a set of events $\boldsymbol{E}$ and $E'(T_2)$ is an instance of a set of events $\boldsymbol{E'}$. Language expresses this by $\texttt{Rel}(E(\cdot), E'(\cdot))$ which can be read as follows: the event $E(\cdot)$ is related to the event $E'(\cdot)$ by the relation $\texttt{Rel}$. In this way, when storing the relationship between the events, language loses the information of the absolute times T_1 and T_2 and accomplishes abstraction of concrete events. This means that it is inherently impossible

 ISBN: 978-1-61208-950-8

for language to restore the lost information about the absolute times. $\texttt{Rel}(E(\cdot), E'(\cdot))$ is stored in the RMD-frame and $E(\cdot)$ and $E'(\cdot)$ are stored in the WMD-frame.

2) Image Can Maintain Reality: The loss of the information concerning absolute times is clearly understood by using a class of words for expressing movement, e.g., *take*, *make*, *get*, and *do*, as an example. They do not hold precise information in the time dimension as they are stored in the WMD-frame.

What happens precisely in the real world when we say "take some flowers to the hospital" is, e.g., 1) buy flowers at 9:00 a.m. at a flower shop; 2) arrive at the hospital at 9:30 a.m. The contents in the WMD-frame relevant to this event are contrasted with those in the PMD-frame. Information stored in the PMD-frame maintains the information about the absolute times. Movement of an object in the real world is defined precisely as a trajectory represented by a time series of the locations of the object in the three-dimensional space. Perceptual information associated with the movement of the object is processed by respective sensory organs at their characteristic sensing rates, e.g., visual information is processed at the rate of 100 msec per characteristic event. The perceptual information may resonate with the contents in the PMD-frame which is an accumulation of the past perceptual experiences for confirming their existence in the PMD-frame.

3) Recovery of Lost Information: The lost information concerning absolute times of events could be recovered with the help of the information stored in the BMD-frame, where the compiled motion patterns for the specific perceptual images are stored without the loss of the information of absolute times, which are critical for producing motion sequences synchronous with the external environment. The BMD-frame is used in System 1 for carrying out image-based inference, which operates at the biological band of the time scale of human action [36, Fig. 3-3] at the time range of < 100 msec.

Imagine you heard the sentence: "My husband brought flowers to the hospital." This verbal information would resonate with the PMD-frame and then activate the event nodes in the WMD-frame that represent relevant events such as "My husband bought flowers," "My husband went to the hospital," and so on, after a series of inferences by applying procedural knowledge stored in the RMD-frame. These nodes are represented as a connected pattern of symbols in the WMD-frame; they are represented by S_i and S_j in Figure 3. The overlap between the WMD- and RMD-frames, and the BMD-frame could activate the nodes in the BMD-frame, which are the compiled motion patterns for the specific perceptual images. S_i and S_j in the WMD-frame could be associated with *Image-i* and *Image-j*, respectively, which might or might not be close to the sight S_i represents. There is no guarantee it is the real image, but the absolute time is recovered with no guarantee of its reality. In this case, the image nodes might represent the motion patterns for the husband's flower-taking and going-to-hospital events, that have been individually encoded at any time in the past.

The use of the information in the BMD-frame has another advantage for the inference processes carried out in System 2. Suppose that *Image-k* follows *Image-j* in the BMD-frame. This could activate the symbol node S_k in the WMD-frame, which results in the establishment of the pattern of connection, $S_i \Rightarrow S_j \Rightarrow S_k$ in the WMD-frame. When the results of inferences are guaranteed by the time-guaranteed BMD-frame, even if there is no guarantee of its reality, they could be used as a shortcut to speed up the inference process. Retrieving images in the BMD-frame through the resonance mechanism would require a lot of effort, which a person would want to avoid. Once these connections are established, they are used as if it is possible to recover the lost information without making an effort to confirm its reality.

4) Reality in Self Experience: Consider another situation where you told a friend, "I took the flowers to the hospital." This describes your own experience and differs from the previous example, which describes the behavior of one's husband. In this case, the utterances heard via your ears would resonate with your PMD-frame followed by spreading activation to the words stored in the WMD-frame, to the procedural rules stored in the RMD-frame, and finally to the encoded sequence of behaviors stored in the BMD- and the MMD-frames, where concrete time series of the behavior just mentioned would be reproduced. In other words, when a person tells his/her own experience to someone else, its reality would be assured in himself or herself; while a person who heard someone else's experience would never reproduce its reality, i.e., recovery of the lost information could not be accomplished completely when language is used for human–human communication.

5) Supporting Reality: How would MHP/RT and the multi-dimensional memory frames function when the input stimuli are images or non-words? As shown in Section II, the multi-dimensional memory frames help the PCM processes map the M-dimensional sensory stimuli onto the N-dimensional motions. The following demonstrates the richness of non-verbal image-based communication compared with verbal word-based communication, which should support the reality.

Let the dimensionality of images be M^{I} and that of words be M^{W}. M^{I} is the number of dimensions for *non-verbal* communication. Let M_i^{I} be the i-th dimension of non-verbal communication and N_i^{I} be the number of discriminable states for M_i^{I}. The number of discriminable states in non-verbal communication, N^{I}, is $\prod_{i=1}^{M^{\mathrm{I}}} N_i^{\mathrm{I}}$. M^{W} is the number of dimensions used for *verbal* communication which corresponds to the number of categories used for representing symbols, such as alphabets, in verbal communication, N^{W}, is $\prod_{j=1}^{M^{\mathrm{W}}} N_j^{\mathrm{W}}$. It would be reasonable to assume that M^{W} is smaller than M^{I}, $M^{\mathrm{W}} < M^{\mathrm{I}}$, and the numbers of discriminable states for any categories of words, M_j^{W}, are smaller than those for any M_i^{I}'s. Therefore, the number of discriminable states differs by an order of magnitude difference in power between the non-verbal space composed of perception and the verbal space composed of the symbols $N^{\mathrm{I}} \ggg N^{\mathrm{W}}$.

6) Seeing is Believing: Finally, consider a situation where no language appears. For example, you feel ephemeral when you are observing the cherry blossoms falling in front of you. The visual event resonates with the memories stored in the PMD-frame, which would activate images in the BMD-frame overlapping with the memory for motor processes stored in the MMD-frame. In other words, the visual event that is occurring in front of you activates the memory traces stored in the BMD- and MMD-frames to help to imagine a variety of situations that could occur next in reality. Symbol nodes in the WMD-frame

 ISBN: 978-1-61208-950-8

would be activated as shown in the upward arrow from the BMD-frame to the RMD- and WMD-frames in Figure 3 for verbally explaining the situation and for deriving the reasons for the event, which leads to an acceptance that the statement is true or that something exists, i.e., the state of believing.

V. Conclusion

This paper discussed the differences between the inference processes initiated by language and image. There are teachings that suggest that image-based inference should be superior to language-based inference, such as "seeing is believing" and "a picture is worth ten thousand words." The experience of seeing or hearing is perceptual. This is transferred to cognitive and motor processes to act appropriately in the external environment. The crucial difference between language and image is the size of the dimensionality when they are perceived. It was suggested that the fact that the dimensionality of image M^{I} is larger than that of language M^{I} should lead to huge differences in the number of discriminable states at the stage of perception. Accordingly, the language input is processed serially in System 2 and the image input is processed in parallel in System 1. The former is a slow, deliberate, and rational process and the latter is a fast, automatic, and instinctive process.

It was suggested that the low expressive power and the slow processing speed in language processing should trade the reality of the represented concept for the rich expression of the concept that includes the information of absolute times attached to the respective concepts. As the information of absolute times is necessary to reproduce the event precisely, representation of events in terms of language is inherently impossible for reproducing the event but it can do so approximately, at best. In the world of language that has lost time, even if reality can almost be restored, human beings tend to use their experience to secure reality and stop making efforts to restore it from the next event onwards.

It was shown that the inherent inability to recover reality is the primary reason for the suggested implication of superiority of images over language in comprehending the situations represented by them from the analysis obtained by applying state-of-the-art cognitive architecture. An image belongs to an individual, which enables him/her to carry out inferences guaranteed by reality. However, words are shared by the public as symbols to communicate one's thoughts. Words might be associated with the patterns of images stored in the BMD-frame by following the memories from the WMD-frame where the words are stored to the BMD-frame via the RMD-frame. In this way, the public symbols in the WMD-frame are indirectly and approximately associated with the individual image patterns in the BMD-frame, which might be individually different. Language-based inference is based on the accumulation of knowledge up to that point, i.e., a meme [15], so it is strongly bound by it and cannot deviate from its scope. Conversely, images form a perception away from language, which makes it easier to approach the truth.

The differences in the image patterns that individuals would associate with a single symbol might be small when they belong to a single culture. This is one of the conditions for words to work as memes [37] where symbols used in a culture are associated with shared image patterns that guarantee reality. In other words, when the same word is used in a communication between individuals from different cultures, there is there is no guarantee that the shared image will correspond with reality. Since a meme is a common interpretation that is valid within the group to which each individual belongs, it differs between groups. The person from one culture who uses the words to communicate his/her thoughts tends to believe that the images associated with the words should be communicated as effectively to the other person irrespective of the culture s/he is from. If the other person is from a different culture, there is little chance of coincidence in the associated images that these two people will infer from the words. The inherent discrepancies in the reality associated with commonly used words will become problematic for globalization.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 20H04290. The author would like to thank Editage (www.editage.com) for English language editing.

References

[1] D. Kahneman, "A perspective on judgment and choice," American Psychologist, vol. 58, no. 9, 2003, pp. 697–720.

[2] J. H. Larkin and H. A. Simon, "Why a diagram is (sometimes) worth ten thousand words," Cognitive Science, vol. 11, no. 1, 1987, pp. 65–100.

[3] J. Hadamard, An Essay on the Psychology of Invention in the Mathematical Field. New York, NY, USA: Dover Publications, 1954.

[4] D. Kahneman, Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux, 2011.

[5] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)," Behaviour & Information Technology, vol. 31, no. 1, 2012, pp. 41–58.

[6] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)," Biologically Inspired Cognitive Architectures, vol. 5, 2013, pp. 82–93.

[7] M. Kitajima and M. Toyota, "Two Minds and Emotion," in COGNITIVE 2015 : The Seventh International Conference on Advanced Cognitive Technologies and Applications, 2015, pp. 8–16.

[8] M. Kitajima, S. Shimizu, and K. T. Nakahira, "Creating memorable experiences in virtual reality: Theory of its processes and preliminary eye-tracking study using omnidirectional movies with audio-guide," in 2017 3rd IEEE International Conference on Cybernetics (CYBCONF), June 2017, pp. 1–8.

[9] M. Kitajima, "Nourishing problem solving skills by performing hci tasks – relationships between the methods of problem solving (retrieval, discovery, or search) and the kinds of acquired problem solving skills," in Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018), vol. 2: HUCAPP. Setúbal, Portugal: SCITEPRESS – Science and Technology Publications, 2018, pp. 132–139.

[10] M. Kitajima, "Cognitive Chrono-Ethnography (CCE): A Behavioral Study Methodology Underpinned by the Cognitive Architecture, MHP/RT," in Proceedings of the 41st Annual Conference of the Cognitive Science Society. Cognitive Science Society, 2019, pp. 55–56.

[11] M. Kitajima, J. Dinet, and M. Toyota, "Multimodal Interactions Viewed as Dual Process on Multi-Dimensional Memory Frames under Weak Synchronization," in COGNITIVE 2019 : The Eleventh International Conference on Advanced Cognitive Technologies and Applications, 2019, pp. 44–51.

[12] J. Dinet and M. Kitajima, "The Concept of Resonance: From Physics to Cognitive Psychology," in COGNITIVE 2020 : The Twelfth International Conference on Advanced Cognitive Technologies and Applications, 2020, pp. 62–67.

 ISBN: 978-1-61208-950-8

[13] J. Dinet and M. Kitajima, "Immersive interfaces for engagement and learning: Cognitive implications," in Proceedings of the 2015 Virtual Reality International Conference, ser. VRIC '18. New York, NY, USA: ACM, 2018, pp. 18/04:1–18/04:8. [Online]. Available: https://doi.org/10.1145/3234253.3234301

[14] M. Kitajima, "Cognitive Science Approach to Achieve SDGs," in COGNITIVE 2020 : The Twelfth International Conference on Advanced Cognitive Technologies and Applications, 2020, pp. 55–61.

[15] M. Kitajima, M. Toyota, and J. Dinet, "The Role of Resonance in the Development and Propagation of Memes," in COGNITIVE 2021 : The Thirteenth International Conference on Advanced Cognitive Technologies and Applications, 2021, pp. 28–36.

[16] M. Kitajima, Memory and Action Selection in Human-Machine Interaction. Wiley-ISTE, 2016.

[17] S. K. Card, T. P. Moran, and A. Newell, The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[18] J. L. McClelland and D. E. Rumelhart, Parallel Distributed Processing: Explorations in the Microstructure of Cognition : Psychological and Biological Models. The MIT Press, 6 1986.

[19] D. Morris, The nature of happiness. London: Little Books Ltd., 2006.

[20] J. L. Bermúdez, Thinking without words. Oxford University Press, 2007.

[21] R. E. Mayer, "Research-based principles for the design of instructional messages: The case of multimedia explanations," Document Design, vol. 1, no. 1, 1999, pp. 7–19. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/dd.1.1.02may

[22] R. E. Mayer, "Multimedia learning," in Psychology of learning and motivation. Elsevier, 2002, vol. 41, pp. 85–139.

[23] A. Paivio, "Dual coding theory and education," in Draft chapter presented at the conference on Pathways to Literacy Achievement for High Poverty Children at The University of Michigan School of Education, 2006.

[24] M. Sadoski and A. Paivio, Imagery and text: A dual coding theory of reading and writing. Routledge, 2013.

[25] V. A. Thompson and A. Paivio, "Memory for pictures and sounds: Independence of auditory and visual codes." Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, vol. 48, no. 3, 1994, p. 380.

[26] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant, "Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies," Cerebral cortex, vol. 19, no. 12, 2009, pp. 2767–2796.

[27] J. Wang, J. A. Conder, D. N. Blitzer, and S. V. Shinkareva, "Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies," Human Brain Mapping, vol. 31, no. 10, 2010, pp. 1459–1468.

[28] R. E. Mayer, "Cognitive theory of multimedia learning," The Cambridge handbook of multimedia learning, vol. 41, 2005, pp. 31–48.

[29] R. Mayer, "Share this page," Journal of Computer Assisted Learning, vol. 33, no. 5, 2017.

[30] D. Mutlu-Bayraktar, V. Cosgun, and T. Altan, "Cognitive load in multimedia learning environments: A systematic review," Computers & Education, vol. 141, 2019, p. 103618.

[31] R. J. Sternberg, Metaphors of mind: Conceptions of the nature of intelligence. Cambridge University Press, 1990.

[32] M. Kejriwal, Q. Wang, H. Li, and L. Wang, "An empirical study of emoji usage on twitter in linguistic and national contexts," Online Social Networks and Media, vol. 24, 2021, p. 100149.

[33] U. Pavalanathan and J. Eisenstein, "Emoticons vs. emojis on twitter: A causal inference approach," arXiv preprint arXiv:1510.08480, 2015.

[34] B. Stiegler. Philosophising by accident: Interviews with elie during (english edition). [retrieved: 4, 2017]

[35] M. C. Corballis, From Hand to Mouth: The Origins of Language. Princeton University Press, 9 2003.

[36] A. Newell, Unified Theories of Cognition (The William James Lectures, 1987). Cambridge, MA: Harvard University Press, 1990.

[37] D. C. Dennett, From Bacteria to Bach and Back: The Evolution of Minds. W W Norton & Co Inc, 2 2018.

 ISBN: 978-1-61208-950-8

No Time to Crash: Visualizing Interdependencies for Optimal Maintenance Scheduling

Milot Gashi
Pro2Future GmbH
Graz, Austria
milot.gashi@pro2future.at

Belgin Mutlu
Pro2Future GmbH
Graz, Austria
belgin.mutlu@pro2future.at

Stefanie Lindstaedt
Graz University of Technology
Graz, Austria
lindstaedt@tugraz.at

Stefan Thalmann
University of Graz
Graz, Austria
stefan.thalmann@uni-graz.at

***Abstract*—With the digital transformation in manufacturing, Predictive Maintenance (PdM) is increasingly proposed as an approach to increase the efficiency of manufacturing processes. However, system complexity increases due to mass customization, shorter product life cycles, and many component variants within a manufacturing system. So far, PdM mainly focuses on a single component or system-level and thus neglects the complexity by not considering the interdependencies between components. In a Multi-Component System (MCS) perspective, models covering interdependencies between components within a complex system are established and used for the prediction. Even if the predictive accuracy is superior, modeling interdependencies is a complex and laborious task that prevents the broad adoption of the MCS perspective. A potential way to tackle this challenge is using visualizations to discover and model the interdependencies. This paper evaluates different visualization approaches for PdM in the context of MCSs using a crowd-sourced study involving 530 participants. In our study, we ranked these approaches based on the participant's performance that aimed to identify the optimal timing for maintenance within an MCS. Our results suggest that visualization approaches are suitable to identify interdependencies and that the stacked-area approach is the most promising approach in this regard.**

***Keywords*—Multi-Component System; Predictive Maintenance; Visualization Analytics; Optimization; Stochastic Interdependencies**

I. INTRODUCTION

Predictive Maintenance (PdM) focuses on managing maintenance actions based on the prediction of component or system conditions. Currently, PdM significantly impacts not only the manufacturing process but also the whole industrial product life cycle [1]. Hence, high-quality products and more reliable manufacturing processes are provided. In practice, PdM is applied either on the level of an entire system or a single component, thus neglecting the interdependencies between components within a system [2]–[4]. However, with the digital transformation in manufacturing, more complex processes, shorter product life cycles, and a wide variety of product variants emerged. This transformation not only increases the complexity of manufacturing systems; it also decreases the time to interact with the system and to learn and understand its behavior. Thus, it is more challenging to model the health indicators accurately and leads to the application of the more simple single component approaches on the one hand. But on the other hand, the grown complexity increases the demand for approaches taking this complexity sufficiently into account, such as Multi-Component System (MCS) view.

MCS models describe interdependencies between components within a system and can be used to improve predictive results and decision support. In the literature, it has been shown that the presence of interdependencies between the components impacts the deterioration process of the components, subsystems, and system [4] [2]. For instance, an old worn-out component will accelerate the wear out of the newly replaced components that interact with it. Therefore, identifying and understanding interdependencies between components helps to extract this additional knowledge, which could contribute to a better understanding of system performance in terms of component and system degradation and improve predictive results. However, this process is not straightforward and requires incorporating human cognitive reasoning and decision-making. Hence, more cognitive approaches applicable to such complex systems with a variety of properties and factors that could influence the decisions are required [5]. In particular, current research within the topic of MCS lacks proper methods to identify interdependencies, thus, failing to build MCS within complex production systems [1]. A promising way to tackle these challenges is through using visualization approaches which provide intuitive and faster ways to understand and identify interdependencies [6].

Visual Analytics (VA) has been applied in PdM for MCSs, aiming to show the presence of interdependencies within an MCS [4] [7] [8]. Nevertheless, the usefulness of visualization aiming to identify interdependencies within an MCS has not been evaluated in these studies. Recently, Gashi et al., [6] evaluated different visualization approaches regarding their suitability for maintenance scheduling. However, to the best of our knowledge, visual approaches with respect to optimal timing were neither analyzed nor ranked so far in the existing research within the field of PdM.

 ISBN: 978-1-61208-950-8

This research work evaluates different visualization approaches based on stochastic interdependencies. In particular, we rank visualization approaches aiming to identify the optimal timing for maintenance, i.e., replace point. Moreover, we discuss difficulties in integrating such approaches in the context of MCS. Additionally, factors that motivate users' decisions for maintenance actions are discussed.

This paper is organised as follows: The second section gives a brief overview of theoretical background. The third section describes materials and methods used for the research experiments. Furthermore, the forth section discusses the results and contribution of this research work, whereas the last section highlights the conclusions and future work.

II. Theoretical Background

In the age of big data, PdM is gaining a lot of attention due to the ability to use predictive models to determine when maintenance actions are required [3]. PdM application provides many benefits, such as decreased maintenance costs and downtime, and increased production performance, sustainability, and quality. In particular, superior predictive results for maintenance help to improve maintenance decisions, such as maintenance scheduling or resource optimization [3]. So far, PdM solutions are acceptable, but in practice, the increased complexity of manufacturing processes and the products leads to the need for more precise results. Moreover, PdM is mainly applied solely on system level or single components, thus neglecting the interdependencies between components completely. In the literature, the presence of interdependencies between components within an MCS can be found [4] [2]. Modeling interdependencies between components could help improve predictive results and understand the deterioration behavior of the components and the system.

In literature, MCS interdependencies are analyzed from different perspectives. Hence, interdependencies are grouped into four different categories: stochastic, economic, structural, and resources based interdependencies. Whereby, stochastic interdependencies analyze the deterioration effect between components within an MCS. For instance, Assaf et al. [4] proposed a deterioration model for MCS, which aims to describe stochastic degradation processes. Economic interdependencies, on the other side, focus on the effect of the costs that can be assured through performed maintenance within an MCS [9]. Structural interdependencies, however, take into account components that are structurally dependent and use this knowledge to improve maintenance processes [10]. Finally, resource-based dependencies aim to model dependency between components and spare parts or other required maintenance resources [11]. In general, modeling and presenting interdependencies within an MCS is a complex and challenging process, which helps to improve predictive results and decision-making processes when performing maintenance. Therefore, presenting and understanding interdependencies to the end-user is a crucial aspect that could help to improve the results. However, the process of identifying interdependencies is not straightforward and requires human cognitive reasoning and decision-making.

Cognitive computing, in general, aims to develop coherent, suitable, adaptable techniques based and inspired in the human mind, which can adapt to new situations [12]. More specifically, cognitive computing is a term used by IBM to describe techniques that can learn from a wide range of datasets, can provide reasons, interact with humans, and learn over time within the context [13]. In particular, understanding and extracting knowledge from big data is an important aspect of handling new emerging data-based decision problems. Therefore, in the context of our work, it is crucial to provide approaches that help facilitate human cognitive reasoning, which could increase shop floor workers' performance and system reliability. One promising approach in this regard is VA.

Visualizations help to understand and extract knowledge from data, thus, improving the decision-making process. Data visualization was applied in various contexts in the manufacturing process, e.g., to identify quality derivations and machine failures in a data-driven way [14], for anomaly detection [15] or for causal analysis [16]. Moreover, visualization for decision making in the context of PdM has been extensively applied [17] [18] [19]. Additionally, extensive research works focusing on PdM for MCSs used visualizations to demonstrate the existence of interdependencies within an MCS [4] [7] [8]. For instance, Assaf et al. [4] used line charts to present the interdependencies between components. Whereby, Shahraki et al. [7] used multi-line to visualize interdependencies within an MCS. Yet, the usefulness of visualization to identify and present interdependencies has not been evaluated.

Nevertheless, Gashi et al. [6], evaluated and ranked visualization approaches for MCSs use-cases in terms of functionality using a crowd-sourced study. In this case, the aim was to identify the best visualization approach that helps users conduct a successful maintenance strategy, i.e., the system will not crash. However, various critical systems, along with the aim to avoid crashes, require to operate at their optimal level. In this case, optimal timing is crucial and a must requirement. To the best of our knowledge, this challenge was not evaluated. Therefore, in this work, we aim to analyze and rank visual approaches in terms of optimal timing for maintenance.

III. Materials and Methods

A. Visualization approaches

Based on a literature review, we first identified candidate visualization approaches for modeling interdependencies: line-based approach [4], matrix-based approach [20] [8], multi-line approach [21], bar-based approach [22], and stacked area approach [23]. Next, in close collaboration and in multiple iterations, we pre-selected the suitable approaches in discussion sessions with three domain experts. As a result, we defined the appropriate rule to pre-select the relevant approaches which are evaluated within this study. A visualization is considered relevant if it fits the following requirements: First, visualization highlights interdependencies over time, emphasizes the

 ISBN: 978-1-61208-950-8

performed maintenance actions, and space reduction, i.e., less space is required for visualization, is an important feature.

Multi-line visualizations are suitable approaches for pattern and relationship analysis of multiple time series data [21]. In a nutshell, deterioration of the components as a physical parameter evolves over the distance and this is shown in multi-line visualization on the x-axis (see Fig. 1). The deterioration rate, on the other side, is shown on the y-axis. Consequently, 1 represents a thoroughly worn-out component and 0 a new component. Furthermore, each line shows the deterioration over distance for a specific component, e.g., chain.

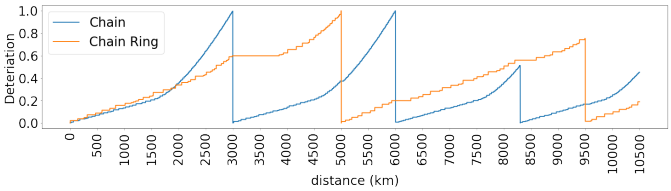


Fig. 1. MCS interdependencies are presented using a multi-line visualization approach. The x-axis represents the distance information in km, and the y-axis represents the deterioration time.

Heatmap approaches are suitable for cross-examination, patterns, or similarity analysis of multivariate data [20]. Heatmap visualizations are built using a matrix format and coloring of cells based on the magnitude of variables. Moreover, space reduction is a crucial feature of this approach. In the heatmap visualization, shown in Fig. 2, the deterioration of the components is encoded visually using a variety of colors for each cell. A cell over the x-axis represents a specific distance ride in km. The color within a specific row represents the corresponding deterioration state of the component. A white color reveals a new component, whereas a black indicates a fully worn-out component. Finally, a row represents a single component's deterioration and maintenance state.

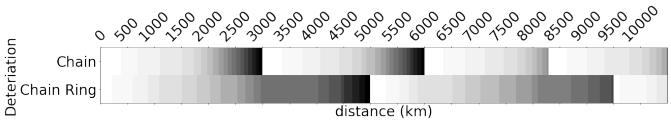


Fig. 2. MCS interdependencies are presented using the heatmap visualization approach.

Stacked-area visualizations (shown in Fig. 3) aim to represent multiple time-series data by stacking filled shapes (single time-series) on top of each other [24]. This approach is relevant for pattern, causal, and comparison analysis. The distance is shown on the x-axis, and the deterioration rate is shown on the y-axis. For each component, the deterioration rate of 1 represents a fully worn-out component, respectively 0 a new component. In contrast to the multi-line approach, the stacked-are approach accumulates the deterioration rate of all components at every specific distance point.

All these approaches can visualize data over time, e.g., over distance, are appropriate for pattern recognition or relationship analysis. Moreover, an advantage of these approaches is the space reduction feature, thus, increasing the relevance of these approaches in use cases, such as MCS, where space is an important aspect due to a large number of components. This work is an extension of previous research work [6]; therefore, further details are explained and published in [6].

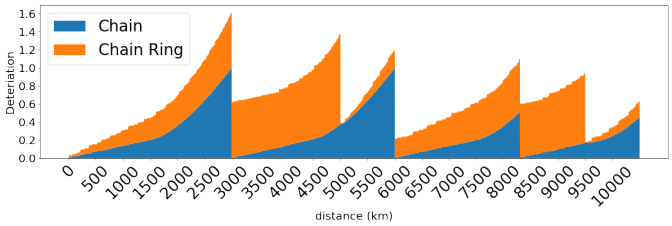


Fig. 3. MCS interdependencies are presented using the stacked-area visualization approach.

B. Procedure and user study

This design study aims to identify which visualization/-s is/are the most appropriate ones for visualizing and identifying the interdependencies of an MCS. For this purpose, a crowd-sourced study is designed to evaluate different visualization approaches. The visualizations are evaluated using a bike example as a common MCS use case that most people are aware of and understand. In particular, the bike has a small number of components, and strong interactions between these components are present; therefore, the bike is an appropriate MCS. In particular, we focus on two specific components based on domain experts' knowledge: chain and chain-ring. Moreover, we used resorted synthetic data to describe the deterioration process and interdependencies between components based on the mathematical model introduced in [6].

In the design study, each participant had to evaluate only one specific visualization in detail, which has been assigned randomly, thus avoiding the presence of biased data [25]. Moreover, the order of answers to all questions within the design study was randomized. As a first step, a description and purpose of the study altogether with the information about confidentiality regarding the data are provided to each participant. Second, the participant is asked to answer some demographic questions, such as expertise on visualization or education level. Further demographic information is collected directly from the platform used to conduct the study. Third, the MCS use case and the definition of interdependencies are presented through short video animation. Next, the participant performed the task to evaluate the assigned visualization approach. This task was designed based on the suggestion from Kittur et al. [26], thus motivating the participants to analyze the visualizations accurately and prevent random answers.

The task was designed as follows: (1) a short description of the visualizations was shown to the participant, (2) two different scenarios of component deterioration over time and performed maintenance actions using the corresponding visualization were shown to the participant. Next, the participant is asked to analyze these scenarios in detail and try to identify the interdependencies between components. As

 ISBN: 978-1-61208-950-8

a next step, the participant is asked to rank the recognized level of interdependencies. Finally, the participant is asked to design a maintenance strategy for a distance ride of 10 000 km having a limited budget of 600 €. Whereby chain costed 20 € and chain-ring 200 €. Next, we performed a usability evaluation based on the System usability scale framework [27]. Moreover, the participant is asked to provide subjective feedback through a post-task questionnaire based on NASA TLX [28] six dimensions of workload: mental demand, physical demand, temporal demand, effort, frustration, and perceived performance. Finally, to each participant, all three approaches are shown, and the participant is asked to select the approach that they would use to identify interdependencies between components.

In total, 704 users participated in the crowd-sourced study. 530 (M=435, F=84, N/A=11) participants, age 18-65 were approved. The approve rule is built based on the strategies provided by participants, thus considering only serious attempts. For instance, strategies containing only random numbers were rejected. As a result, 72 submissions were rejected. Moreover, 89 users returned their submission, i.e., they interrupted the participation and did not submit the result. Finally, 13 participation were rejected from the platform for exceeding the time limit, which was 45 minutes (m) by default. The participants had experience with visual- and data analytic tools. Moreover, all participants had experience in the industry and were well educated (529 participants with at least a bachelor's degree). Participants needed 12 m and 42 seconds (s) on average to analyze the heatmap approach successfully. Participants successfully analyzed stacked-area visualization in 13 m and 42 s on average and the multi-line approach in 12 m and 50 s on average.

C. Evaluation of maintenance strategies

For the analysis, all valid strategies (non-crash strategies) are considered and evaluated based on the mathematical model introduced in [6]. To estimate the optimal timing (replace-point detection) for maintenance, we estimated the average deviation to the optimal replace-point for every maintenance replacement entry with respect to provided strategies. In this case, the best point is 0, which indicates no deviation; respectively, maintenance was performed at optimal timing. Consequently, positive deviation indicates that participants replaced the component after the optimal timing exceeded, thus indicating that the optimal timing was not recognized. Negative deviation indicates that the participants performed maintenance before the optimal timing was reached. Bar plots with confidence intervals are used to visually quantify and evaluate the results. This helps to easily identify and compare the distributions' mean and the confidence intervals. Moreover, the non-parametric Mann-Whitney U test [29] was used as a statistical approach to estimate the difference between distributions. Results with p-value < 0.05 are considered as significant differences.

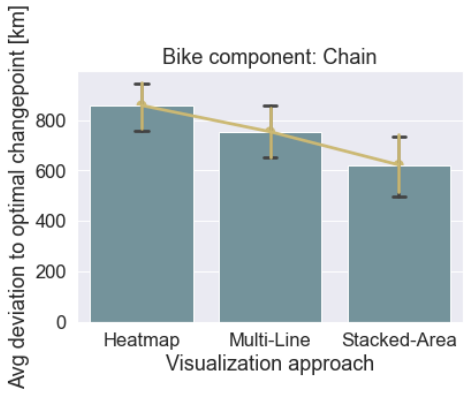


Fig. 4. Bike Chain: Estimated average deviation to change-point (optimal timing) overall participants concerning each visualization approach.

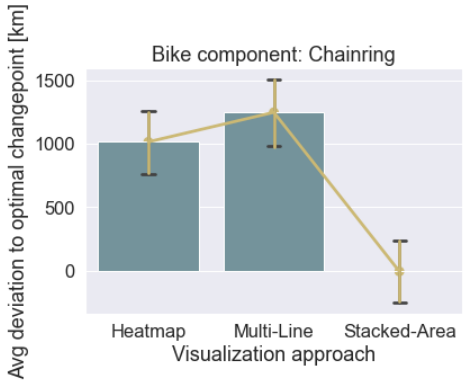


Fig. 5. Bike Chainring: Estimated average deviation to change-point (optimal timing) overall participants concerning each visualization approach.

IV. Discussion of Results

In Fig. 4 and 5, results concerning optimal timing for both chain and chain-ring are shown. Whereby 0 represents the optimal deterioration time. The positive deviation indicates a defensive approach (replacement performed after optimal timing), respectively, negative deviation indicates a more offensive approach (replacements performed before optimal timing). In this case, all three approaches are compared with each other. As a result, the multi-line and heatmap approach show similar behavior with respect to both components, i.e., chain and chain-ring. Consequently, the stacked-are approach outperforms both approaches significantly, see confidence intervals. Moreover, the Mann-Whitney U test results demonstrate the significance of this result with $p < 0.05$. Furthermore, every participant who analyzed a visualization approach in this study was asked to provide qualitative feedback regarding the identification level of interdependencies they manage to identify as shown in Fig. 6.

The participant ranked the approaches between 0 and 5, where 0 indicates that no interdependencies were identified and 5 that strong interdependencies were obvious from the corresponding approach. Correspondingly, participants who

 ISBN: 978-1-61208-950-8

analyzed the stacked-area approach seem more confident that they were able to identify interdependencies. These results are statistically significant. In general, the results demonstrate that the stacked-area visualization approach significantly outperforms the other visualization approaches with respect to optimal timing. In this regard, stacked-area visualization is a more offensive approach compared to the other approaches. Still, it is not clear what contributes to these early decisions, and more in-depth research is needed. But one possible factor could be the accumulated deterioration degree showed within a stacked-area approach. This could trigger early decisions and thus reaction on time. The stacked-area approach as an offensive approach could be appropriate in sensitive settings, where breakdown should be prevented due to safety or cost demand. However, in the previous study [6], stacked-area approaches showed the highest error rate (around 44%) with respect to strategies that lead the system to crash. This could be related to the background knowledge level of the participants or the required training; however, there were no significant results from this study in this regard. In particular, the accumulated deterioration state shown as staked areas on top of each other (multiple components) within a staked-area approach could increase the distortion effects while interpreting results within stacked-area visualization [30], thus, leading to a higher error rate. Thus, our study delivered the first evidence that visualization approaches could be used to identify interdependencies in the context of PdM. We found that the visualization approaches perform differently. More research seems promising to identify the suitability of the different visualization approaches for different PdM settings and MCS modeling approaches.

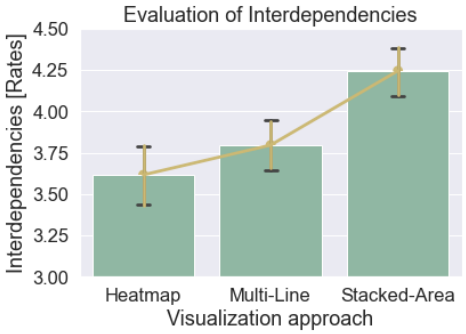


Fig. 6. The results of subjective feedback with respect to interdependencies are shown.

Predominantly, the evaluation of visualization approaches for complex systems, such as MCSs is not a trivial process due to the many factors influencing the process and users' decisions. Factors, such as domain expertise on maintenance or visualization approaches could lead to different results. The results within this study with respect to the background of participants show a trend, where the more experienced the users are, the better they perform; however, the results are not significant, and yet further research in this regard remains feasible. Moreover, depending on the goal of business concerning the sensitivity of the manufacturing process that requires maintenance, different approaches might be suitable. For instance, Gashi et al., [6] showed that the Multi-line is a suitable approach to perform maintenance that avoids downtime but does not necessarily perform at the optimal time. In contrast, in this work, we showed that the stacked-area approach is appropriate when aiming for maintenance at optimal timing. This leads to the conclusion that different perspectives potentially lead to different results. As Plaisant [31] suggests, it requires studying and manipulating data repetitively from multiple perspectives over a long time in order to discover new knowledge. Similarly, Roberts [32] encourages the analysis of data from multiple views while using visualization approaches in order to avoid false conclusions or misinformation. Therefore, researchers could be encouraged to consider these results as a further avenue for future research. In general, this research work demonstrated that simple visualizations could identify the interdependencies concerning optimal timing. In the future, we plan to explore more complex visualizations and xAI approaches in terms of VA, which seem promising in this regard.

V. Conclusion

This research work showed that visualization approaches are suitable to identify interdependencies in the context of PdM. Our key finding bases on a design study to analyze and rank visualization approaches involving 530 participants. The stacked-area approach turned out to be the best approach in terms of optimal timing, thus, being a relevant approach in more sensitive cases, where downtime should be avoided due to safety or cost reasons. Finally, we discussed that the context and business goals within a complex MCS impact the selection of the appropriate visualization approach and that more research is needed to inform the selection of visualization approaches.

In this design study, participants had to generate strategies for the short term, i.e., 10 000 km; therefore, in the future, it will be interesting to compare these approaches in the long term (longer than 10 000 km). Moreover, user interviews in such a study could help understand user behavior regarding maintenance decisions. Furthermore, we evaluated visualization approaches using a simple MCS containing only two components. In the future, an evaluation of these approaches against a complex MCS (larger number of components) is required. Furthermore, synthetic data are used to model the scenarios for all three approaches; therefore, in the future, evaluating these approaches using data from a real use case could provide new insights. For this purpose, we plan to integrate these approaches within a DSS and evaluate them in a real industrial use case.

Acknowledgment

This work has been supported by the FFG, Contract No. 881844: Pro^{2}Future is funded within the Austrian COMET Program Competence Centers for Excellent Technologies under the auspices of the Austrian Federal Ministry for Cli-

mate Action, Environment, Energy, Mobility, Innovation and Technology, the Austrian Federal Ministry for Digital and Economic Affairs and of the Provinces of Upper Austria and Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

REFERENCES

[1] S. Thalmann, H. G. Gursch, J. Suschnigg, M. Gashi *et al.*, "Cognitive decision support for industrial product life cycles: A position paper," in *Proceedings of the 11 th International Conference on Advanced Cognitive Technologies and Applications*, 2019, pp. 3–9.

[2] L. Bian and N. Gebraeel, "Stochastic framework for partially degradation systems with continuous component degradation-rate-interactions," vol. 61, no. 4. Wiley Online Library, 2014, pp. 286–303.

[3] M. Gashi and S. Thalmann, "Taking complexity into account: A structured literature review on multi-component systems in the context of predictive maintenance," in *European, Mediterranean, and Middle Eastern Conference on Information Systems*. Springer, 2019, pp. 31–44.

[4] R. Assaf, P. Do, P. Scarf, and S. Nefti-Meziani, "Wear rate-state interaction modelling for a multi-component system: Models and an experimental platform," vol. 49, no. 28. Elsevier, 2016, pp. 232–237.

[5] S. Thalmann, J. Mangler, T. Schreck *et al.*, "Data analytics for industrial process improvement a vision paper," in *2018 IEEE 20th Conference on Business Informatics (CBI)*, vol. 2. IEEE, 2018, pp. 92–96.

[6] M. Gashi, B. Mutlu, S. Lindstaedt, and S. Thalmann, "Decision support for multi-component systems: visualizing interdependencies for predictive maintenance," in *Hawaii International Conference on System Sciences 2022 (HICSS 2022)*, 2022.

[7] A. F. Shahraki, A. Roy, O. P. Yadav, and A. P. S. Rathore, "Predicting remaining useful life based on instance-based learning," in *2019 Annual Reliability and Maintainability Symposium (RAMS)*. IEEE, 2019, pp. 1–6.

[8] M. C. O. Keizer, R. H. Teunter, and J. Veldman, "Joint condition-based maintenance and inventory optimization for systems with multiple components," vol. 257, no. 1. Elsevier, 2017, pp. 209–222.

[9] K.-A. Nguyen, P. Do, and A. Grall, "Joint predictive maintenance and inventory strategy for multi-component systems using birnbaum's structural importance," vol. 168. Elsevier, 2017, pp. 249–261.

[10] A. Van Horenbeek and L. Pintelon, "A dynamic predictive maintenance policy for complex multi-component systems," vol. 120. Elsevier, 2013, pp. 39–50.

[11] M. Gashi, P. Ofner, H. Ennsbrunner, and S. Thalmann, "Dealing with missing usage data in defect prediction: A case study of a welding supplier," vol. 132. Elsevier, 2021, p. 103505.

[12] D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, "Cognitive computing," vol. 54, no. 8. ACM New York, NY, USA, 2011, pp. 62–71.

[13] M. Mohammadi and A. Al-Fuqaha, "Enabling cognitive smart cities using big data and machine learning: Approaches and challenges," vol. 56, no. 2. IEEE, 2018, pp. 94–101.

[14] M. Gashi, B. Mutlu, J. Suschnigg, P. Ofner, S. Pichler, and T. Schreck, "Interactive visual exploration of defect prediction in industrial setting through explainable models based on shap values," in *IEEE VIS Poster Program*, 2020.

[15] H. Janetzko, F. Stoffel, S. Mittelstädt, and D. A. Keim, "Anomaly detection for visual analytics of power consumption data," vol. 38. Elsevier, 2014, pp. 27–37.

[16] X. Xie, F. Du, and Y. Wu, "A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications," vol. 27, no. 2. IEEE, 2020, pp. 1448–1458.

[17] A. Cachada, J. Barbosa, P. Leitño, Gcraldcs *et al.*, "Maintenance 4.0: Intelligent and predictive maintenance system architecture," in *2018 IEEE 23rd international conference on emerging technologies and factory automation (ETFA)*, vol. 1. IEEE, 2018, pp. 139–146.

[18] J. C. Cheng, W. Chen, K. Chen, and Q. Wang, "Data-driven predictive maintenance planning framework for mep components based on bim and iot using machine learning algorithms," vol. 112. Elsevier, 2020, p. 103087.

[19] S.-j. Wu, N. Gebraeel, M. A. Lawley, and Y. Yih, "A neural network integrated decision support system for condition-based optimal predictive maintenance policy," vol. 37, no. 2. IEEE, 2007, pp. 226–236.

[20] L. Wilkinson and M. Friendly, "The history of the cluster heat map," vol. 63, no. 2. Taylor & Francis, 2009, pp. 179–184.

[21] R. J. Pandolfi, D. B. Allan, E. Arenholz, Barroso-Luque *et al.*, "Xi-cam: a versatile interface for data visualization and analysis," vol. 25, no. 4. International Union of Crystallography, 2018, pp. 1261–1270.

[22] F. Chang, G. Zhou, C. Zhang, Z. Xiao, and C. Wang, "A service-oriented dynamic multi-level maintenance grouping strategy based on prediction information of multi-component systems," vol. 53. Elsevier, 2019, pp. 49–61.

[23] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, "Fluid interaction for information visualization," vol. 10, no. 4. Sage Publications Sage UK: London, England, 2011, pp. 327–340.

[24] A. Thudt, J. Walny, C. Perin, F. Rajabiyazdi *et al.*, "Assessing the readability of stacked graphs," in *Proceedings of Graphics Interface Conference (GI)*, 2016, pp. 167–174.

[25] Y. Weinstein and H. L. Roediger, "Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it," vol. 38, no. 3. Springer, 2010, pp. 366–376.

[26] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2008, pp. 453–456.

[27] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," vol. 4, no. 3. Citeseer, 2009, pp. 114–123.

[28] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[29] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other." JSTOR, 1947, pp. 50–60.

[30] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," vol. 79, no. 387. Taylor & Francis Group, 1984, pp. 531–554.

[31] C. Plaisant, "The challenge of information visualization evaluation," in *Proceedings of the working conference on Advanced visual interfaces*, 2004, pp. 109–116.

[32] J. C. Roberts, "On encouraging multiple views for visualization," in *Proceedings. 1998 IEEE Conference on Information Visualization. An International Conference on Computer Visualization and Graphics (Cat. No. 98TB100246)*. IEEE, 1998, pp. 8–14.

 ISBN: 978-1-61208-950-8

Reasoning and Arguments in Negotiation

Developing a formal model

Mare Koit
Institute of Computer Science
University of Tartu
Tartu, Estonia
e-mail: mare.koit@ut.ee

***Abstract*—Our aim is to develop a model of negotiation where two participants present arguments for/against of doing an action. The choice of an argument depends, on one hand, on the beliefs about the positive and negative aspects of doing the action and the needed resources, and on the other hand, on the result of reasoning affected by these beliefs. The model is based on the analysis of human-human negotiations (in this paper, we consider telemarketing calls). A limited version of the model is implemented as a dialogue system. The computer attempts to influence the reasoning of the user by its arguments in order to convince the user to make a decision.**

Keywords—reasoning; beliefs; negotiation; argument; dialogue system.

I. INTRODUCTION

Negotiation is a form of interaction in which a group of agents, with a desire to cooperate but with potentially conflicting interests try to come to a mutually acceptable division of scarce resources [1]. Negotiation is simultaneously a linguistic and a reasoning problem, in which intent must be formulated and then verbally realized. A variety of agents have been created to negotiate with people within a large spectrum of settings including the number of parties, the number of interactions, and the number of issues to be negotiated [2]. Negotiation dialogues contain both cooperative and adversarial elements, and their modelling require agents to understand, plan, and generate utterances to achieve their goals [3].

Our aim is to develop a model of conversational agent that interacts with a user in Estonian and carries out negotiation. We start with the analysis of human-human negotiations aiming to model the reasoning processes which people go through when pursuing their communicative goals and coming to a decision.

The rest of the paper is organized as follows. Section 2 describes related work. In Section 3, we analyze a kind of human-human negotiation dialogues – telemarketing calls, in order to explain how do people reason and argue when negotiating about doing an action. In Section 4, we introduce our model of conversational agent that takes into account the results of the analysis of human-human negotiations, and an implementation – a simple Dialogue System (DS). Section 5 discusses the model and the DS. In Section 6, we draw conclusions and plan future work.

II. RELATED WORK

A conversational agent, or DS, is a computer system intended to interact with a human using text, speech, graphics, gestures and other modes for communication. It will have both dialogue modelling and dialogue management components [4]. A dialogue manager is a component of a DS that controls the conversation. Four kinds of dialogue management architectures are most common: plan-based, finite-state, frame-based, and information-state [5]. An information state includes beliefs, assumptions, expectations, goals, preferences and other attitudes of a dialogue participant that may influence the participant's interpretation and generation of communicative behavior. The functions of the dialogue manager can be formalized in terms of information state update [6].

Rahwan et al. [7] discuss three approaches to automated negotiation: game-theoretic, heuristic-based and argumentation-based. Argumentation-based approaches to negotiation allow agents to 'argue' about their beliefs and other mental attitudes during the negotiation process. Argumentation-based negotiation is the process of decision-making through the exchange of arguments [3].

In negotiation, an argument can be considered as a piece of information that may allow an agent to: (a) justify its negotiation state; or (b) influence another agent's negotiation state [8]. Amgoud and Cayrol define an argument as a pair (*H, h*) where: (i) *H* is a consistent subset of the knowledge base, (ii) *H* implies *h*, (iii) *H* is minimal, so that no subset of *H* satisfying both (i) and (ii) exists. *H* is called the support and *h* the conclusion of the argument [9].

Automated negotiation agents capable of negotiating efficiently with people must rely on a good opponent modelling component to model their counterpart, adapt their behavior to their partner, influencing the partner's opinions and beliefs [10]. NegoChat is the first negotiation agent successfully developed to use a natural chat interface while considering its impact on the agent's negotiation strategy [2]. A virtual human that negotiating with a human helps people learn negotiation skills. For virtual agents, the expression of attitudes in groups is a key element to improve the social believability of the virtual worlds that they populate as well as the user's experience, for example in entertainment or training applications [11][12][13].

 ISBN: 978-1-61208-950-8

An interesting and useful kind of DSs are embodied conversational agents [11][14][15].

III. Analysis of Human-Human Negotiations

Our further aim is to implement a DS which interacts with a user in Estonian and carries out negotiations like a human does. For that, we are studying human-human negotiations using the Estonian dialogue corpus [16]. All the dialogues are recorded in authentic situations and then transliterated by using the transcription of Conversation Analysis [17]. A sub-corpus of telemarketing calls is chosen for the current study. In the dialogues, two official persons are communicating – a sales clerk of an educational company (its changed name is Tiritamm, he is the initiator of a call), and a manager or a personnel officer of another institution (she is here a customer). Tiritamm offers training courses (management, sale, etc.) which can be useful for the employees of the customer's institution. The communicative goal of a sales clerk is to convince the customer to decide to take a course. Several typical phases can be differentiated in telemarketing negotiations [18]. The most important phase is *argumentation*. A sales clerk (*A*) presents different arguments that take into account the actual needs of the customer (*B*) explained by him (*A*) before or during the call. *A* tries to bring out the factors that are essential for the customer, in order to convince her to make a positive decision (Example 1). If *B* accepts these factors then *A* will demonstrate/prove how the proposed course will solve *B*'s problem. In an ideal case, the customer will agree with the proof offered by the clerk and will decide to take the course.

Example 1 (transcription of Conversation Analysis is used in the examples)

A: /---/(1.0) .hh sest loomu`likult et=ee `töökogemuste kaudu: õpib ka: alati aga .hh a `sageli ongi just `see (0.5) mt ee `kursused pakuvad sellise `võimaluse kus saab siis `teiste .hh oma hh `ala `spetsia`listidega samuti `kokku=ja `rääkida nendest `ühistest prob`leemidest ja samas siis ka .hh ee `mõtteid ja `ideid ee hh ee=`Tiritamme poolt sinna `juurde.

because, of course, one can learn from experience but frequently training courses make it possible to meet other specialists in the field and discuss common problems; additional thoughts and ideas come from Tiritamm argument

(.)

B: £ `jah?

yes accept

The behaviour of sales clerks and customers is different when they are arguing for/against a course. A sales clerk when having the initiative provides his arguments for taking the course either asserting something (then a customer typically accepts the assertion, Example 1). A customer, to the contrary, does not accept assertions/arguments of a sales clerk when arguing against taking a course (Example 2).

Example 2

/---/ *B*: aga jah ei mul on see läbi ´vaadatud=ja (.) ´kahjuks ma pean ütlema=et (.) et ´teie (.) seda meile (.) ´ei suuda ´õpetada (.) mida (.)´mina: (.) tahan.

but I have looked through your catalogue of courses and unfortunately, I have to say that you can't teach what is needed for us counter argument

/---/

A: .h ja mida kon´kreetselt=ee ´teie tahate.

and what do you want question

(0.8) mida te ´silmas ´peate.

what do you have in view question

B: noo (0.2) ´meie (.) äri´tegevus on (.) ´ehitamine.

well, our business is house-building answer

/---/

A: nüüd kas (0.2) näiteks (0.5) ´lepingute ´saamisel (0.5) mt ee ´tegelete te ka: läbi´rääkimistega.

well, do you need to carry out negotiations in order to achieve agreements question

B: noo ikka.

yes, of course answer

(0.8)

A: mt et see=on ka üks ´valdkond mida me: (0.2) ´käsitleme.

but that is one of our fields which we cover argument

The argumentation continues in a similar way. *A* attempts to convince *B* preparing his new arguments by questions. Either *A* constructs his arguments during the conversation or he chooses suitable arguments from an existing set of possible arguments collected by previous experience with customers.

When modelling negotiation, a good way seems to follow the sales clerks' strategy: try to take and hold the initiative and propose 'hard' arguments for the strived action, i.e., the statements that do not provoke the partner's rejection but accept. In order to have such arguments at disposal, it is necessary to know as possible more about the partner in relation to the goal action.

IV. Modelling Conversational Agent

Results of the analysis of human-human negotiations motivated our model of conversational agent. Let us consider negotiation between two participants *A* and *B* where *A* is the initiator. Let his communicative goal be to bring *B* to the decision to do an action *D*. When convincing *B*, he is using a partner model (a picture of the communication partner) that gives him grounds to believe that *B* will agree to do the action. *A* starts the dialogue by proposing *B* to do *D*. If *B*, as the result of her reasoning, refuses, then *A* must influence her in the following negotiation, continuously correcting the partner model and trying to guess in which reasoning step *B* reached her negative decision. In this way, a dialogue – a sequence of utterances will be generated together by *A* and *B*.

A. Information States in Negotiation Process

Let *A* and *B* be conversational agents. Let us assume the following [18]:

1) a set *G* of communicative goals where both participants choose their own initial goals (G^A and G^B, respectively). In our case, G^A = "*B* decides to do *D*"

2) a set *S* of communicative strategies of the participants. A communicative strategy is an algorithm used by a participant for achieving his/her communicative

goal. This algorithm determines the activity of the participant at each communicative step

3) a set *T* of communicative tactics, i.e., methods of influencing the partner when applying a communicative strategy. For example, *A* can entice, persuade, or threaten *B* in order to achieve the goal G^A, i.e., *A* attempts to demonstrate that achieving this goal is, accordingly, pleasant, useful or obligatory for *B*

4) a set *R* of reasoning models, which are used by participants when reasoning (here: about doing an action *D*). A reasoning model is an algorithm, which returns the positive or negative decision about the reasoning object (the action *D*)

5) a set *P* of participant models, i.e., a participant's depiction of the beliefs of himself/herself and his/her partner in relation to the reasoning object: $P = \{P^A(A), P^A(B), P^B(A), P^B(B)\}$

6) a set of world knowledge

7) a set of linguistic knowledge.

A conversational agent passes several *information states* during interaction starting from initial state and going to every next state by applying *update rules*. Information states represent cumulative additions from previous actions in the dialogue, motivating future actions. There are two parts of an information state of a conversational agent [7] – private (information accessible only for the agent) and shared (accessible for both participants).

Two categories of *update rules* will be used by a conversational agent for moving from current information state into the next one: (1) for interpreting the partner's turns and (2) for generating its own turns.

B. Reasoning Model

The initial version of our reasoning model is introduced in [19]. In general, it follows the ideas realized in the BDI (Belief-Desire-Intention) model [20]. The reasoning process of a subject about doing an action *D* consists of steps where the resources, positive and negative aspects of *D* will be weighed. A communication partner can only implicitly take part in this process by presenting arguments to stress the positive and downgrade the negative aspects of *D*.

Our reasoning model includes two parts: (1) a model of (human) motivational sphere that represents the beliefs of a reasoning subject in relation to the aspects of the action under consideration, and (2) reasoning procedures.

1) Model of Motivational Sphere

We represent the model of *motivational sphere* of a communication participant as a vector with (here: numerical) coordinates that express the beliefs of the participant in relation to different aspects of the action *D*:

w_D = *(w(resources$_D$), w(pleasant$_D$), w(unpleasant$_D$), w(useful$_D$), w(harmful$_D$), w(obligatory$_D$), w(prohibited$_D$), w(punishment-do$_D$), w(punishment-not$_D$))*.

The value of *w(resources$_D$)* is 1 if the reasoning subject has all the resources needed for doing *D*, and 0 if some of them are missing. The value of *w(obligatory$_D$)* or *w(prohibited$_D$)* is 1 if the action is obligatory or, respectively, prohibited for the subject (otherwise 0). The values of the other coordinates can be numbers on the scale from 0 to 10 – *w(pleasant$_D$), w(unpleasant$_D$)*, etc, indicate the values of the pleasantness, unpleasantness, etc. of *D* or its consequences; *w(punishment-do$_D$)* is the punishment for doing a prohibited action and *w(punishment-not$_D$)* – the punishment for not doing an obligatory action.

2) Reasoning Procedures

The reasoning itself depends on the *determinant*, which triggers it. With respect to the used theory, there are three kinds of determinants that can cause humans to reason about an action *D*: wish, need and obligation [21]. Therefore, three different prototypical *reasoning procedures* can be described – WISH, NEEDED, and MUST. Every procedure consists of steps passed by a reasoning subject and it finishes with a decision: do *D* or not. When reasoning, the subject considers his/her resources as well as different positive and negative aspects of doing *D*. If the positive aspects (pleasantness, usefulness, etc.) weigh more than negative (unpleasantness, harmfulness, etc.) then the decision will be "do *D*" otherwise "do not do *D*". The reasoning subject checks primarily his/her wish, thereafter need and then obligation and he/she triggers the corresponding reasoning procedures. If no one procedure returns the decision "do *D*" then the reasoning ends with the decision "do not do *D*".

In Figure 1, we present the reasoning procedure NEEDED, which is triggered by the need of the reasoning subject to do the action *D* (i.e., doing the action is more useful than harmful for the subject) The procedure is presented as a step-form algorithm. We do not more indicate the action *D*.

Presumption: w(useful) ≥ w(harmful).
1) Is w(resources) = 1? If not then go to 8.
2) Is w(pleasant) > w(unpleasant)? If not then go to 5.
3) Is w(prohibited) = 1? If not then go to 7.
4) Is w(pleasant) + w(useful) > w(unpleasant) + w(harmful) + w(punishment-do)? If yes then go to 7 otherwise go to 8.
5) Is w(obligatory) = 1? If not then go to 8.
6) Is w(pleasant) + w(useful) + w(punishment-not) > w(unpleasant) + w(harmful)? if not then go to 8.
7) Decide: do D. End.
8) Decide: do not do D.

Figure 1. Reasoning procedure NEEDED.

We use two vectors w^B and w^{AB}, which capture the beliefs of communication participants in relation to the action *D* under consideration. Here w^B is the model of motivational sphere of *B* who has to make a decision about doing *D*; the vector includes *B*'s (actual) evaluations (beliefs) of *D*'s aspects. These values are used by *B* when reasoning about doing *D*. The other vector w^{AB} is the partner model that includes *A*'s beliefs concerning *B*'s beliefs in relation to the action. It is used by *A* when planning next turns in dialogue. We suppose that *A* has some preliminary knowledge about *B* in order to compose

 ISBN: 978-1-61208-950-8

the initial partner model before making the initial proposal.

Both the models will change as influenced by the arguments presented by both the participants in negotiation. For example, every argument presented by *A* targeting the usefulness of *D* will increase the corresponding values of w^B*(useful)* as well as w^{AB}*(useful)*.

C. Implementation

A simple dialogue system is developed that carries out negotiations with a user in a natural language about doing an action [18]. The participants can have different initial goals: the initiator (either DS or a user) tries to achieve the decision of the partner to do the action but the partner's goal can be opposite. DS interacts with a user using texts in a natural language. There are two work modes. In one case, the computer is playing *A*'s and in the other – *B*'s role.

Both *A* and *B* have access to a common set of reasoning procedures. They also use fixed sets of dialogue acts and the corresponding utterances in a natural language, which are pre-classified semantically, e.g., the set $P_{missing\_resources}$ for indicating that some resources for doing a certain action *D* are missing (e.g., *I don't have proper dresses*, see example 3 in the next section), $P_{increasing\_resources}$ for indicating that there exist resources for (e.g., *The company will cover all your expenses*), $P_{increasing\_usefulness}$ for stressing the usefulness of *D* (e.g., *You can be useful for the company*), etc. Therefore, no linguistic analysis or generation will be made during a dialogue. The utterances will be accidentally chosen by conversational agent from the suitable semantic classes (in our implementation, every utterance can be used only once). However, these restrictions will bring along that the generated dialogues are not quite coherent.

If *A*'s goal is "*B* will do *D*" and *B*'s goal is opposite then *A*, starting interaction, generates, by using his knowledge a partner model w^{AB} and determines the communicative tactics *T*, which he will use (e.g., persuasion), i.e., he accordingly fixes the reasoning procedure *R*, which he will try to trigger in *B*'s mind (e.g., NEEDED). *B* has her own model w^B (which exact values *A* does not know). *B* in her turn determines a reasoning procedure R^B that she will use in order to make a decision about doing *D*, and her communicative tactics T^B.

D. Reasoning and Arguments

When attempting to direct *B*'s reasoning to the positive decision (do *D*), *A* presents several arguments stressing the positive and downgrading the negative aspects of *D*. Preparing an argument, *A* triggers a reasoning procedure in his partner model w^{AB}, in order to be sure that the reasoning will give the positive decision. When opposing, *B* can use the same or a different reasoning procedure triggering it in the model of herself w^B. After the changes made by both the participants in the two models during a dialogue, the models will approach each to other but, in general, do not equalise. Although, the results of reasoning in both models can be (or not be) equal.

Let us consider a dialogue with our DS (Example 3). Here *A* is the conversational agent playing the role of the boss of a company and *B* is the user playing the role of an employee of the company who is at the same time studying at a university. *A* presents arguments for doing *D* by *B* (*D*=travel to *N.* in order to conclude a contract there). It succeeds to decline *B*'s counter arguments and convince *B* to accept its goal.

Example 3

```
    1. A: The company offers you a trip to N.
Our company needs to conclude a contract there.
    2. B: I don't have proper dresses.
    3. A: The company will pay your executive
expenses. You can be useful for the company.
    4. B: I can have some problems at my
university.
    5. A: It's all right–your examinations
period will be extended. The company will
evaluate your contribution.
    6. B: OK, I'll do it.
    7. A: I am glad.
```

Let us consider how the partner model is used in the dialogue. *A* will implement the tactics of *persuasion* and generates a partner model, let it be

w^{AB}={w^{AB}*(resources)=1*, w^{AB}*(pleasant)=4*, w^{AB}*(unpleasant)=2*, w^{AB}*(useful)=5*, w^{AB}*(harmful)=2*, w^{AB}*(obligatory)=0*, w^{AB}*(prohibited)=0*, w^{AB}*(punishment-do)=0*, w^{AB}*(punishment-not)=0*}.

The reasoning procedure NEEDED (Figure 1) yields a positive decision in this model. *A*'s *initial information state* is as follows.

Private part

- initial partner model w^{AB} = (1, 4, 2, 5, 2, 0, 0, 0, 0)
- the tactics chosen by *A*–persuasion
- *A* will use the reasoning procedure NEEDED, the presumption is fulfilled: w^{AB}*(useful)* > w^{AB}*(harmful)*
- the set of dialogue acts at *A*'s disposal: {proposal, arguments for increasing/decreasing values of different coordinates of w^{AB}, accept, reject}
- the set of utterances for expressing the dialogue acts at *A*'s disposal: {*The company offers you a trip to N*, *You can be useful for the company*, etc.}.

Shared part

- the reasoning procedures WISH, NEEDED, MUST
- the tactics of enticement, persuasion, threatening
- dialogue history–an empty set.

Let us suppose that every statement (argument) presented in dialogue will increase or respectively, decrease the corresponding value in the model of beliefs by one unit. Still, this is a simplification because different arguments might have different weights for different dialogue participants.

Conversational agent *A* starts the dialogue with a proposal. Using the tactics of persuasion and attempting to trigger the reasoning procedure NEEDED in *B*, it adds an argument for increasing the usefulness to the proposal (turn 1). In the same time, it increases the initial value of the usefulness in its partner model w^{AB} by 1. The current reasoning procedure NEEDED still gives a positive

 ISBN: 978-1-61208-950-8

decision in the updated model. *A* does not know the actual values of attitudes, which *B* has assigned in the model w^B of herself. As caused by every counter argument presented by *B*, *A* has to update the partner model w^{AB}. However, *B*'s counter argument (turn 2) demonstrates that *B* actually has resources missing (*I don't have proper dresses*) therefore, *A* has to decrease the value of w^{AB}*(resources)* from 1 to 0 in its partner model. Now *A* must find an argument indicating that the resources are available: it selects an utterance from the set $P_{increasing\_resources}$ (*The company will pay your executive expenses*) and following the tactics of persuasion it adds an argument for increasing the usefulness (*You can be useful for the company*) in turn 3. The value of w^{AB}(*resources)* will now be 1 and the value of w^{AB}*(useful)* will be increased by 1 in the partner model. The reasoning in the updated model gives a positive decision. Nevertheless, *B* has a new counter argument indicating the harmfulness of the action: *I can have some problems at my university* (turn 4).

Now *A* has to increase the value w^{AB}*(harmful)* in the partner model, it turns out that by 6 not by 1 as we assumed. Let us explain why. So far, *A* was supposing that *D* is not prohibited for *B*. This assumption proves to be wrong because otherwise it is impossible for *B* to indicate the harmfulness of *D* (if she is applying the reasoning procedure NEEDED as *A* supposes). Therefore, *B* supposedly compares the values of beliefs at the step 4 of the procedure and makes a negative decision. *B* can come to the step 4 only after the step 3 where she detects that *D* is prohibited and doing *D* involves a punishment (turn 4). Therefore, *A* changes the value of w^{AB}*(prohibited)* from 0 to 1 and increases the value of w^{AB}*(punishment-do)* in the partner model at least by 1. (Being optimistic, *A* increases the value exactly by 1 and not more.) Now *A* checks, how to change the value of the harmfulness in the partner model in order to get the negative decision like *B* did. According to the reasoning procedure NEEDED *A* calculates that the value has to be increased (at least) by 6. Therefore, w^{AB}*(harmful)* will be 2+6=8.

Responding to *B*'s counter argument *A* decreases the value of w^{AB}*(harmful)* by 1 using the utterance *It's all right - your examinations period will be extended*, and increases the value of w^{AB}*(useful)* once more using the utterance *The company will evaluate your contribution* (turn 5). The reasoning procedure NEEDED gives a positive decision in the updated partner model. Now it turns out that *B* has made this same decision (turn 6). *A* has achieved its communicative goal and finishes the dialogue (turn 7).

Example 3 demonstrates how *A* is updating the partner model w^{AB} in negotiation with *B*. The final model will be w^{AB} = (1, 4, 2, 8, 7, 0, 1, 0, 1). As compared with the initial model, the values of four aspects have increased. All the changes are caused by *A*'s arguments and *B*'s counter arguments.

In this way, *A* is able to convince *B* to do *D* if he has enough arguments for doing *D* and his initial picture of *B* does not radically differentiate from *B*'s actual beliefs. Both the beliefs in the partner model w^{AB} and *B*'s actual beliefs in the model w^B of herself (if *B* is a conversational agent similarly with *A*) are changing during the dialogue as influenced by the arguments presented by the participants. Although the models w^{AB} and w^B do not necessarily coincide at the end of the dialogue, the proportions of the values of the positive (pleasantness, etc.) and negative aspects of doing *D* (unpleasantness, etc.) will be similar. Still, if *B* is a human user then she is not obliged to use models and algorithms (although can) but can choose utterances from suitable semantic classes.

V. DISCUSSION

Our model of conversational agent is motivated by the analysis of human-human negotiations. We consider the dialogues where two participants *A* and *B* negotiate doing an action *D*. In the analysed telemarketing calls, the communicative goal of a sales clerk of the educational company is to convince a customer to take a training course offered by the company. If the participants are collaborative and one of them presents his/her argument then the partner mostly accepts it. If the participants are antagonistic then at least one of them does not agree with the opinion of the partner and presents his/her counterargument(s). The more the clerk knows about the customer, the more convincing arguments is he able to choose. Asking questions is a way to learn more.

When reasoning about doing an action, a subject is weighing different aspects of the action (its pleasantness, usefulness, etc.), which are included into his/her model of motivational sphere. In the model presented here, we evaluate these aspects by giving them discrete numerical values on the scale from 0 to 10. Still, people do not use numbers but rather words of a natural language, e.g., *excellent, very pleasant, harm*, etc. Further, when reasoning, people do not operate with exact values of the aspects of an action but they rather make 'fuzzy calculations', for example, they suppose/believe that doing an action is more pleasant than unpleasant and therefore they wish to do it. Another problem is that the aspects of actions considered here are not fully independent. For example, harmful consequences of an action as a rule are unpleasant. In addition, if the reasoning object is different (not doing an action like in our case) then the attitudes of a reasoning subject can be characterized by a different set of aspects.

When attempting to direct *B*'s reasoning to the desirable decision, *A* presents several arguments stressing the positive and downgrading the negative aspects of *D*. The choice of *A*'s argument is based on one hand, on the partner model, which captures *A*'s knowledge about *B*, and on the other hand, on the (counter) argument presented by *B*. Still, *B* is not obliged to present any counter argument but she can only refuse (*I do not do this action*). When choosing the next argument for *D*, *A* triggers a reasoning procedure in his partner model depending on the chosen communicative tactics, in order to be sure that the reasoning will give a positive decision after presenting this argument. *B* herself can use the same or a different reasoning procedure triggering it in her own model. After the updates made both by *A* and *B* in the two models during a dialogue, the models will approach each

 ISBN: 978-1-61208-950-8

to another but, in general, do not equalize. Nevertheless, the results of reasoning in both models can be similar, as demonstrated example 3. Therefore, *A* can convince *B* to do *D* even if not having a perfect picture of her.

Our dialogue model considers only a limited kind of dialogues but although, it illustrates the situation where the dialogue participants are able to change their beliefs related to the negotiation object and bring them closer one to another by using arguments. The initiator *A* does not need to know whether the counter arguments presented by the partner *B* have been caused by *B*'s opposite initial goal or are there simply obstacles before their common goal, which can be eliminated by *A*'s arguments. *A*'s goal, on the contrary is not hidden from *B*. Secondly, the different communicative tactics used by *A* are aimed to trigger different reasoning procedures in *B*'s mind. *A* can fail to trigger the pursued reasoning procedure in *B* but however he can achieve his communicative goal when having a sufficient number of arguments supporting his initial goal.

In our implemented DS, the user interacts with the computer, choosing ready-made, semantically pre-classified sentences as arguments and counter arguments for and against doing a certain action. We suppose that this kind of software is useful when training the skills of finding arguments and counter arguments for and against of doing an action. The computer can establish certain restrictions on the argument types and on the order in their use. Still, when interacting with the computer, a human user does not use neither a formal partner model, nor a formal model of herself, nor reasoning procedures. However, both implementation modes allow study how the beliefs of the participants are changing in negotiation.

VI. Conclusion and Future Work

We are considering the dialogues where two (human or artificial) agents *A* and *B* negotiate doing an action *D* by one of them (*B*). We analyse human negotiations in order to explain how arguments are used to convince a dialogue partner. Initial communicative goals of the participants can be similar or opposite. The partners present arguments for and against of doing *D*. The arguments of initiator *A* are based on his partner model w^{AB} whilst *B*'s arguments – on her model of herself w^{B}. Both models include beliefs about the resources, positive and negative aspects of doing *D* that have numerical values in our implementation. Both models are updated during a dialogue.

Our further aim is to develop the DS concentrating foremost on the reasoning model. So far, we are using an intuitive reasoning theory. However, there are several other approaches to model change of a person's opinion, e.g., Elaboration Likelihood Model, Social Judgment Theory, and Social Impact Theory. Some of the theories can be better to model human reasoning. Our further research will explain this.

Acknowledgment

This work was supported by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies).

References

[1] I. Rahwan, P. McBurney, and L. Sonenberg, "Towards a theory of negotiation strategy (a preliminary report)," Game-Theoretic and Decision-Theoretic Agents (GTDT), S. Parsons and P. Gmytrasiewicz, Eds. Proc. of an AAMAS-2003 Workshop, pp. 73–80, 2003. Melbourne, Australia.

[2] A. Rosenfeld, I. Zuckerman, E. Segal-Halevi, O. Drein, and S. Kraus, "NegoChat: A Chat-based Negotiation Agent," Proc. of the International Conference on Autonomous Agents and Multi-agent Systems AAMAS'14, pp. 525–532, 2014.

[3] M. Lewis, D. Yarats, Y.N. Dauphin, D. Parikh, and D. Batra, "Deal or No Deal? End-to-End Learning for Negotiation Dialogues," Proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 2443–2453, 2017. Copenhagen, Denmark, Assoc. for Computational Linguistics.

[4] D. Traum, Computational Approaches to Dialogue, *The Routledge Handbook of Language and Dialogue*, 2017.

[5] D. Jurafsky and J. M. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2009.

[6] D. Traum and S. Larsson, "The Information State Approach to Dialogue Management," *Current and New Directions in Discourse and Dialogue*, pp. 325–353, 2003.

[7] I. Rahwan et al., "Argumentation-based negotiation", The Knowledge Engineering Review, vol. 18(4), pp. 343–375, Cambridge University Press, 2004. Available from http://dx.doi.org/10.1017/S0269888904000098

[8] N. R. Jennings, S. Parsons, P. Noriega, and C. Sierra, "On argumentation-based negotiation," Proc. of the International Workshop on Multi-Agent Systems, Boston, pp. 1–7, 1998.

[9] L. Amgoud and C. Cayrol, "A Reasoning Model Based on the Production of Acceptable Arguments", Ann. Math. Artif. Intell. 34(1-3), pp. 197–215, 2002.

[10] C Hadjinikolis, Y Siantos, S Modgil, E Black, and P. McBurney, "Opponent modelling in persuasion dialogues," Proc. of the 23rd IJCAI, pp. 164-170, 2013.

[11] B. Ravenet, A. Cafaro, B .Biancardi, M. Ochs, and C. Pelachaud, "Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions", *IVA*, pp. 375–388, 2015.

[12] J. Gratch, S. Hill, L.-P. Morency, D. Pynadath, and D. Traum, "Exploring the Implications of Virtual Human Research for Human-Robot Teams," VAMR'15, International Conference on Virtual, Augmented and Mixed Reality, R. Shumaker and S. Lackey, Eds. Springer, pp. 186–196, 2015, doi: 10.1007/978-3-319-21067-4_20

[13] M. Saberi, S. DiPaola, and U. Bernardet, "Expressing Personality Through Non-verbal Behaviour in Real-Time Interaction," Frontiers in Psychology, vol 12, pp. 54–74, 2021.

[14] S. Dermouche, "Computational Model for Interpersonal Attitude Expression", *ICMI*, pp. 554–558, 2016.

[15] K. Jokinen, "Exploring Boundaries among Interactive Robots and Humans", Conversational Dialogue Systems for the Next Decade, L. F. D'Haro, Z. Callejas, and S. Nakamura, Eds. Springer: Singapore, LNEE, vol. 704, pp. 271-275, 2020.

[16] T. Hennoste et al, "From Human Communication to Intelligent User Interfaces: Corpora of Spoken Estonian," Proc. of LREC'08, European Language Resources Association (ELRA), pp. 2025–2032, 2008, Marrakech, Morocco. Available from www.lrec-conf.org/proceedings/¬lrec2008

[17] J. Sidnell, Conversation Analysis: An Introduction, London: Wiley-Blackwell, 2010.

[18] M. Koit, "Reasoning and communicative strategies in a model of argument-based negotiation," *Journal of Information and Telecommunication TJIT*, Taylor & Francis Online, 2, 14 p. 2018.

[19] M. Koit and H. Õim, "A Computational Model of Argumentation in Agreement Negotiation Processes", *Argument & Computation*, Taylor & Francis Online, 5 (2-3), pp. 209–236, 2014.

[20] M. E. Bratman, "Intention, Plans, and Practical Reason," *CSLI Publications*, 1999.

[21] H. Õim, "Naïve Theories and Communicative Competence: Reasoning in Communication," *Estonian in the Changing World*, pp. 211–231, 1996.

 ISBN: 978-1-61208-950-8

Guidelines for Designing Interactions

Between Autonomous Artificial Systems and Human Beings

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

Jérôme Dinet
Université de Lorraine, CNRS, INRIA, Loria
Nancy, France
Email: jerome.dinet@univ-lorraine.fr

***Abstract*—Human beings live in an environment that consists of various artifacts, such as physical or virtual tools, information systems, and social systems. With IT advancement, the wider the network of artifacts, the more autonomous they become. However, the ultimate goal of developing these artifacts is to achieve the Sustainable Development Goals (SDGs) through the exploration of the design space for realizing a sustainable society. The artifacts that human beings interact with apply this mechanism for utilizing the artifacts, by selecting the subsequent actions for a given situation. This mechanism includes Perceptual, Cognitive, and Motor (PCM) processes and the memory process. The cognitive process is characterized by the bounded rationality and by the satisficing principle proposed by Simon, and Two Minds of unconscious and conscious processes proposed by Kahneman. The state-of-the art cognitive architecture, Model Human Processor with Realtime Constraints (MHP/RT), developed by Kitajima and Toyota, defines these processes as autonomous systems and proposes a resonance mechanism between the PCM and memory processes. The purpose of this study is to propose guidelines to conduct strategical explorations in design space. Based on the simulation of human–artifact interaction processes through the MHP/RT cognitive architecture, the guidelines are grouped into three levels: goal, mode, and process levels. A method for applying the proposed guidelines while exploring the design space is also presented.**

Keywords–Design guidelines, Human–artifact interaction; Autonomous systems; Cognitive architecture; MHP/RT.

I. INTRODUCTION

When viewed as an individual, each human being is composed of autonomous systems that control perception, cognition, and movement in synchronization with changes in the environment, in addition to a memory autonomous system that works to link perceived movements [1][2]. The environment in which humans interact and live is composed of various artifacts. With the progress of networking technology, a large number of artifacts have become related to each other, overcoming the physical constraints of time and space. In this case, the central management method of the set of artifacts and the environment design to achieve their goals is not effective. It would rather be effective to design the environment as a set of autonomously operating artifacts equipped with Parallel Distributed Processing (PDP), which can be referred to as the Artificial PDP (A-PDP) system, and to design them so that they function as a whole and achieve their goals.

There exist interfaces between the above-mentioned autonomous systems, which have to be properly designed for well-being. The interfaces and internal algorithms defining their behaviors must support activities conducted by human beings; they attempt to achieve their happiness goals by utilizing artifacts. A research question that arises is – how can such interfaces and internal algorithms be designed for the two autonomous systems? Our daily life is based on interactions with a wide variety of artifacts. The purpose of interactions, for human-beings, is to achieve well-being through activities in domains such as health (e.g., bio-monitoring), mobility (e.g., driving an electric vehicle), education (e.g., learning on Massive Open Online Course (MOOC)), and entertainment (e.g., playing e-sports). The artifacts support human activities through the interface at each moment of interaction. There are multiple autonomous systems on both sides of the interface with complex relationships. The purpose of this study is to propose a set of guidelines that should be applied when designing the interfaces of autonomous artifacts, for supporting activities carried out by autonomous human beings.

The remainder of this paper is organized as follows. Section II outlines the human-artifact interaction to define the specific perspective for considering the complex situation of interaction, i.e., both sides are autonomous systems. Section III briefly reviews the Model Human Processor with Realtime Constraints (MHP/RT), developed by Kitajima and Toyota [1][2] and defines a framework for developing the guidelines. A set of guidelines are described. Section IV presents a hint for applying the guidelines to determine the direction of strategic development for the maximum utilization of the artifact. Section V concludes the paper by summarizing the specific role of the guidelines for realizing a sustainable society.

II. HUMAN–ARTIFACT INTERACTION

There are human beings on this side of the interface and artifacts on the other side. From the viewpoint of a user that perceives the information provided on an interface to select the next action for accomplishing his/her goal, a complete understanding of the detailed processing of an artifact to generate information on the interface, e.g., the knowledge of implementing the internal algorithms, is unnecessary; similarly, a detailed understanding of the internal processing of an input to an artifact is unnecessary for them to continue the interaction cycle of execution and evaluation. Although the internal processes are not known to the user, s/he has to comprehend the mechanism at the interface level in order to proceed, i.e., "bridging the gulf between execution and evaluation [3, Figure 3.2]." This also applies to the artifacts. For designers to specify the interfaces of the I/O for the

 ISBN: 978-1-61208-950-8

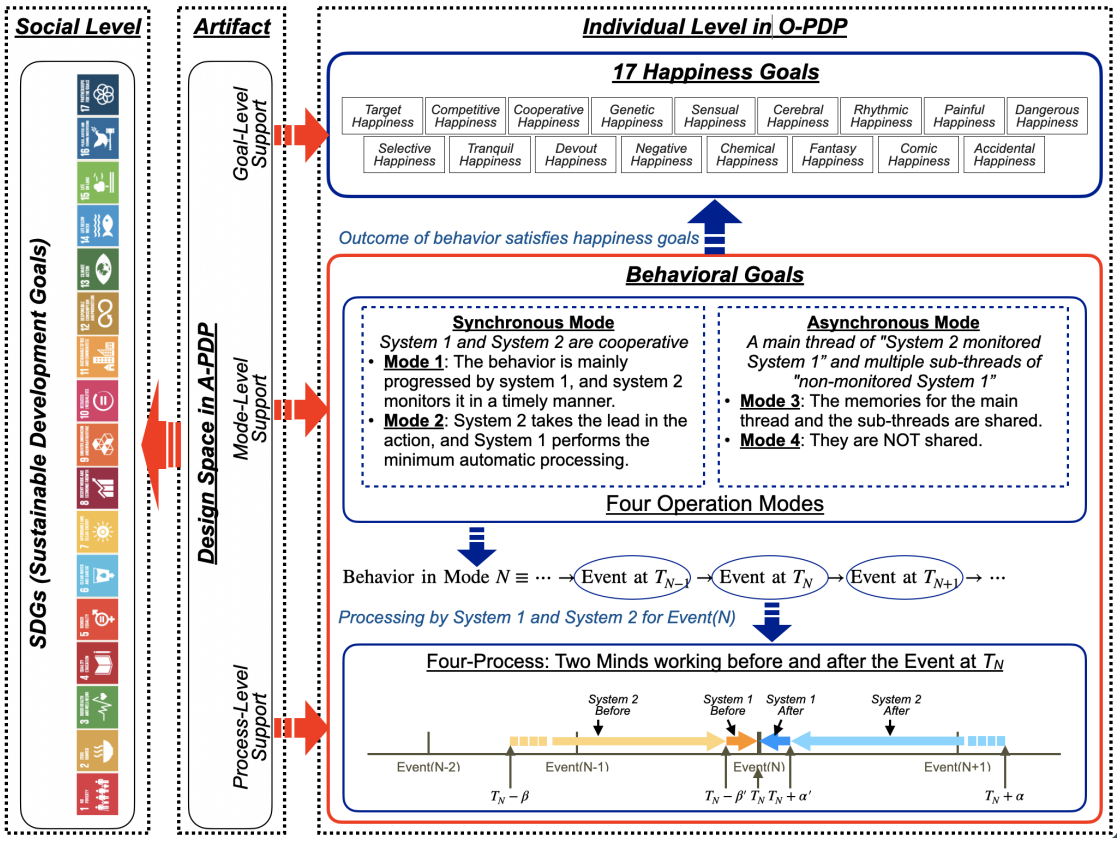


Figure 1. Top-level view of human–artifact interaction.

systems by developing internal algorithms to support human activities, there is no need for them to have a complete understanding of human reactions to the output of the artifacts and of human expectations attached to the input to the artifacts.

As Simon pointed out [4], an interface is characterized by an artificial system between two environments–inner and outer, i.e., human beings and artifacts, respectively. These environments lie in the province of "natural science" where the systems of artifacts and human beings are the focus of research, but the interface linking them is the realm of "artificial science." Therefore, the research question that this study addresses is in the realm of artificial science. The two sides, i.e., the behaviors of human beings and artifacts, are governed by their own principles, and they have to interact with each other by simultaneously considering the behaviors of the either side at *the appropriate approximation levels* in hope of a successful development. Their articulation could be formalized as guidelines, which is the form of an answer to the research question that this study addresses.

The interface between the two systems can be conceived from a variety of perspectives or dimensions. One of them is the dimension that focuses on the Perceptual, Cognitive, and Motor (PCM) processes and the manner in which memory is acquired, used, and developed in the use of artifacts. This study specifically focuses on the ongoing PCM processes and the manner in which they use the memory in the human–artifact interaction process. Our previous study [5] focused on the acquisition and development process and proposed guidelines for designing artifacts, which could cause the evolution of artifacts. In the process of evolution, the *techniques* used in the development of artifacts are received and absorbed by users as *skills* by applying the PCM and memory processes, which is simulated by MHP/RT. The techniques could turn into skills if the conditions derived by the simulations MHP/RT are satisfied. When this spiral evolution occurs, the socio-cultural ecology, wherein the artifacts are embedded, evolves to exhibit a splicing evolution. The focus of this paper is not on the evolution that occurs at the interfaces but on the ongoing events.

Another dimension, which this study effectively focuses on, is the structure of human goals, which can be used by human beings to organize their behaviors. Our previous paper [6] proposed an effective method for achieving Sustainable Development Goals (SDGs) through the behavior of individual human beings, by applying the knowledge of cognitive science; the idea is to connect the daily activities of human beings when trying to accomplish task goals through real world problem-solving [7], i.e., activities in the COGNITIVE Band of Newell's time scale of human action [8, page 122, Fig. 3-3], through any of the SDGs, that concerns social ecology and resides in the SOCIAL Band by finding the non-linear mappings between the goals in different bands. The interfacing situation this study deals with is analogous to the one above. Each individual human being conducts activities to accomplish his or her behavioral goal. This activity is non-linearly mapped onto the autonomous artifacts which have the goal of any of

 ISBN: 978-1-61208-950-8

the SDGs, where the gulfs of execution and evaluation have to be bridged.

III. Interaction Levels and Guidelines

Figure 1 shows the top-level view of the human–artifact interaction. The artifacts placed at the center should exist as entities for achieving any of the SDGs by providing appropriate support for the individual human beings who try to achieve any of the seventeen happiness goals. This section begins by introducing MHP/RT [1][2] in Section III-A focusing on the levels of interactions with artifacts. It follows Section III-B and Section III-C with suggestions for enabling conditions that artifacts have to satisfy to help human beings achieve a smooth coordination between System 1 and System 2. Section III-D presents the relationships between the happiness goals of human beings and the SDGs that the artifacts are expected to achieve. The top-level constraint for developing guidelines is that any artifact that complies with the guidelines has to provide a stable human-artifact interaction; unstable interactions should result in unpredictable results, which do not come with the SDGs.

A. MHP/RT and Interaction Levels

Kitajima and Toyota [1][2] constructed a comprehensive theory of action selection and memory, MHP/RT, that provides a basis for constructing any model to understand the daily behavior of human beings. MHP/RT is an extension of the Model Human Processor (MHP) proposed by Card, Moran, and Newell [9], which can simulate routine goal-directed behaviors. MHP/RT extends the MHP by the following assumptions to consider the fact that the processes involved in action selection are a dynamic interaction that evolves in the irreversible time dimension:

1) The fundamental processing mechanism of the brain is PDP [10], which leads to a collection of the autonomous systems having specific functions for generating an organized human behavior. It consists of autonomous systems for perception, cognition, motor-control, and memory, which is referred to as the Organic PDP (O-PDP) system in the development of MHP/RT.
2) Human behavior emerges as a result of the cooperation of the dual processes of System 1, i.e., fast unconscious processes for intuitive reaction with feedforward control, which connect perception with motor movements, and System 2, i.e., slow conscious processes for deliberate reasoning with feedback control. System 1 and System 2 are referred to as Two Minds [11].
3) Human behavior is organized under 17 happiness goals [12].

Human beings use artifacts to accomplish certain behavioral goals for realizing the desired state of affairs. The human–artifacts interaction is a cycle of PCM processes. The MHP/RT simulates the PCM processes as follows. The cognitive process is to select the next actions that are appropriate for accomplishing the behavioral goals, given the comprehension results of the perceived information. System 1 directly connects to the motor process, whereas System 2 can only indirectly affect the motor process via System 1. The MHP/RT assumes a resonance mechanism for connecting the PCM processes and memory, where the records of the results of the PCM processes are accumulated in a layered and partially overlapped structure of multidimensional memory frames. The cognitive process is carried out by coordinating System 1 and System 2 appropriately to accomplish the behavioral goals. System 1 and System 2 interact simultaneously with the multidimensional memory frames to select an appropriate action and carry it out in a timely manner in the ever-changing environment. The former is the issue of coordination, while the latter is that of synchronization. Section III-B and Section III-C, will address these issues.

TABLE I. Four operation modes of MHP/RT.

Synchronous Modes
- Mode 1: Unconscious mechanism driven mode *A single set of perceptual stimuli initiates feedforward processes at the BIOLOGICAL and COGNITIVE bands to act with occasional feedback from an upper band, i.e., COGNITIVE, RATIONAL, or SOCIAL.*
- Mode 2: Conscious mechanism driven mode *A single set of perceptual stimuli initiates a feedback process at the COGNITIVE band, and upon completion of the conscious action selection, the unconscious automatic feedforward process is activated at the BIOLOGICAL and COGNITIVE bands for action.*
Asynchronous Modes
- Mode 3: In-phase autonomous activity mode *A set of perceptual stimuli initiates feedforward processes at the BIOLOGICAL and COGNITIVE bands with one and another intertwined occasional feedback process from an upper band, i.e., COGNITIVE, RATIONAL, or SOCIAL.*
- Mode 4: Heterophasic autonomous activity mode *Multiple threads of perceptual stimuli initiate respective feedforward processes at the BIOLOGICAL and COGNITIVE bands, some with no feedback and others with feedback from the upper bands, i.e., COGNITIVE, RATIONAL, or SOCIAL.*

B. Mode LevelMode Level: Coordination of Two Minds According to the Goals

Individual beings interact with artifacts to accomplish their behavioral goals by selecting appropriate actions, by running a cycle of PCM processes. The MHP/RT assumes that the action selection processes are controlled by System 1 and System 2. System 1 and System 2 cooperate to connect perception with motion, and the degree of cooperation varies depending on the external environmental conditions, i.e., the state of the artifact that the MHP/RT is interacting with.

1) Four Operation Modes: The conduction of the cooperation can be understood by observing the interaction processes for a certain amount of time, to identify the feature of the

interaction in terms of the mode. The processes carried out by System 1 and System 2 are independent for some time durations but are totally dependent on each other in other domains. This provides a macroscopic view of the manner in which the human-artifact interaction is organized.

Four qualitatively different modes are identified [13]. System 1 is a fast feedforward control process with the characteristic time range of ¡150 ms to connect the perceptual process with the motor process, which makes it possible to behave synchronously with the ever-changing environment. There could be multiple System 1 processes that correspond to active perceptual–motor controls. However, System 2 is a slow feedback control process, which takes a significantly longer time. The time range can be months or years as long as feedback from the past event could affect the ongoing processing. System 2 is a serial process. It can process only one thing at a time; the process could be monitoring one of the active threads of System 1 to check for possible deviations of the results of System 1 from the expected course of actions.

Table I lists four modes, each of which is characterized by the relationships between System 1 and System 2. Modes 1 and 2 are characterized by a single major System 1 process monitored by System 2. The differences between them is the degree of intervention of System 2 for checking the output of System 1. In Mode 1, the occasional feedback from System 2 is sufficient to conduct the behavior. In Mode 2, a frequent monitoring is necessary to organize the behavior appropriately in the environment. Mode 3 corresponds to the situation wherein a single set of perceptual stimuli initiates System 1 processes with one and another intertwined occasional feedback processes by System 2. Mode 4 corresponds to the situation where multiple threads of perceptual stimuli initiate the corresponding System 1 processes, some with no feedback and others with feedback from System 2.

2) Guidelines for Supporting Mode Level Interactions: The human–artifact interaction is carried out in one of the four operation modes of MHP/RT. For the viewpoint of a sound human–artifact interaction, the artifacts should support the interactions that are carried out in Mode 1 and Mode 2. Mode 3 and Mode 4 include unmonitored feedforward System 1 processes, which might cause an instability in the human–artifact interaction. The safety of the human–artifact interaction is realized by allowing the artifact to intervene through System 2, causing the human being to restore to the normal interaction. In other words, the resilience of the human–artifact interaction is realized by maintaining the interaction in Mode 1 and Mode 2. Achieving resilience is a necessary condition for the sustainability of the artifacts to achieve the SDGs.

a) Supporting Mode 2 Interaction: In Mode 2, System 2 frequently intervenes the PCM processes conducted by System 1. More precisely, the pace of interaction with the artifact is controlled by System 2. The role of System 1 is to carry out the necessary PCM processes, to advance the main System 2–artifact interactions. Because System 2 operates on language, the appropriate input from the artifact by means of language is of critical importance. Because the processes of System 1 are carried out in the context defined by those of System 2, the appropriate interactions from the artifact for supporting the processes of System 1 have to be provided considering the context.

Guideline [A]

1. *Converse with System 2.*
2. *Intervene System 1 for facilitating the main conversation with System 2.*

b) Supporting Mode 1 Interaction: In Mode 1, where the intervention of System 2 is weak, language is not an appropriate medium for communication. The interaction from the artifact has to support the unconsciously carried out automatic processes by System 1. However, in Mode 1, the timely examination of the progress is critical for a smooth interaction. The triggers for initiating the examinations carried out by System 2 could be provided internally or externally, i.e., from the artifact. There could be a situation where the examination by System 2 has not been carried out when necessary. In this situation, the intervention from the artifact is necessary for maintaining Mode 1 interaction.

Guideline [B]

1. *For a normal Mode 1 interaction, provide information to both System 1 and System 2, so that System 1-led processes can run smoothly.*
2. *For an intensive Mode 1 interaction, e.g., video games and e-sports, focus on System 1 support.*

c) Supporting Transition Between Mode 1 and Mode 2: When the interaction running in Mode 1 breaks down, it becomes impossible to continue. In this case, the accomplishment of the goal via the interaction being advanced is either given up or a remedial action is taken to return from the failed state to the original normal state and resume to the execution in Mode 1. Mode 2 addresses the recovery process.

Guideline [C]

1. *On the detection of the intensive behavior of System 2 during Mode 1 support, switch from Mode 1 support to Mode 2 support.*

C. Process Level: Synchronization of Two Minds with the Environment

The mode-level support described in Section III-B is defined for the interactions that span the extended time along the time dimension. Therefore, its basis is a macroscopic bird's-eye view of the interactions. However, process-level support is defined for each event that occurs along the time dimension. Its basis is a microscopic view for the interaction at the level of each PCM process. The MHP/RT defines four processing modes by considering the manner in which System 1 and System 2 concern the event occurring at time T.

 ISBN: 978-1-61208-950-8

1) Four Processing Modes: Conscious/Unconscious Processes Before/After an Event: Experiences associated with the activities of an individual are characterized by a series of events, each of which is recognized by a person consciously. As shown in Figure 1, the behavior is defined as a time series of events, "$\cdots \rightarrow$ [Event at T_{N-1}] $\rightarrow$ [Event at T_N] $\rightarrow$ [Event at T_{N+1}] $\cdots$." We focus on *a particular event* that occurs at the absolute time T_N. For the event to occur at T_N, the MHP/RT assumes that there should have existed the conscious processes of System 2 and unconscious processes of System 1 before T_N. For the executed event at T_N, the MHP/RT assumes that there might exist unconscious System 1 processes and conscious System 2 processes, concerning the event after T_N. The behavior of the MHP/RT appears as though it works in one of four processing modes [1] [14] at a time before and after the event at T_N. They are shown at the bottom of Figure 1.

Two of the four processing modes concern the processes carried out *before* the event:

- *System 2 Before Mode:* In the time range of $T_N - \beta \leq t < T_N - \beta'$, where $\beta' \sim 500$ms and β rranges a few seconds to hours or even to months, the MHP/RT uses a part of the memory for System 2 to *consciously* prepare for future events.
- *System 1 Before Mode:* In the time range of $T_N - \beta' \leq t < T_N$, the MHP/RT *unconsciously* coordinates motor activities to the interacting environment. This mode uses the part of the memory for System 1.

The other two modes concern the processes carried out *after* the event:

- *System 1 After Mode:* In the time range of $T_N < t \leq T_N + \alpha'$, where $\alpha' \sim 500$ms, the MHP/RT *unconsciously* tunes the connections between the sensory inputs and motor outputs for a better performance for the same event in the future. This mode updates the connections within the part of the memory for System 1.
- *System 2 After Mode:* In the time range of $T_N + \alpha' < t \leq T_N + \alpha$, the MHP/RT *consciously* recognizes an event in the past and then modifies the memory concerning the event, where α ranges a few seconds to minutes or even to hours. This mode modifies the connections of the part of the memory for System 2.

2) Guidelines for Supporting Process Level Interactions: The human–artifact interaction needs to be synchronized for the cyclic PCM processes to run smoothly in any mode, i.e., Mode 1–4 defined in Section III-B, the interaction is in. The synchronization between the artifact and user is discussed in [15] in the case of a multi-modal interaction using the concepts of four processing modes. "Synchronization" and its derived concept of "weak synchronization" are defined as follows [16]:

> $\cdots$a system and a user is synchronized if every system event at T_{sys} occurs as a user event at T_{user} with some amount of time allowance of Δ, $|T_{\text{user}} - T_{\text{sys}}| < \Delta$, where the actual values of Δ depend on the nature of interactions.
> $\cdots$
> However, a person's activity related with an event has to be considered from the four processing modes, which ranges relatively long time before and after the actual time the event happens. Therefore, "synchronization" has to be considered alternatively as the phenomena a person's activities during the time range of $[T-\beta, T+\alpha]$, which are linked with the specific recognizable system event at time T through a sequence of processes carried out in either of the four processing modes: all the processes have some link with the system event at T. When this is satisfied, the event is considered synchronized with a person's activities, which is called *weak synchronization* [15].

The human–artifact interaction has to provide a smooth flow of the four processing modes. It can break when a person has to adjust his/her activity while s/he is in the *System 1 Before Mode* in such a way that his/her movement should be in synchrony with the current environment. This is the situation that the interaction has to avoid. This is because when this happens, the condition for weak synchronization is not satisfied. To remedy this, s/he has to make extra efforts to re-establish a weak synchronization by adjusting his/her movement. This leads to the following guidelines.

Guideline [D]

1. *Provide appropriate language-level support for System 2 while the user is in System 2 Before Mode, $T_N - \beta \leq t < T_N - \beta'$.*
2. *Provide appropriate perceptual- and motor-level support for System 1 while the user is in System 1 Before Mode, $T_N - \beta' \leq t < T_N$.*

D. Goal Level

The mode-level support described in Section III-B and the process-level support described in Section III-C concern direct interactions with the environment, to accomplish behavioral goals in problem-solving activities, e.g., real-world problem solving [7], or routine goal-oriented skilled activities by applying well-organized knowledge of Goals, Operators, Methods, and Selection rules (GOMS) [9]. As shown in Figure 1, the MHP/RT assumes that the behavioral goals are subordinate to happiness goals; the accomplishment of the behavioral goals are likely to be accompanied by the unconscious feeling of happiness, i.e., achieving a certain happiness goal.

1) Happiness Goals and their Relationship with the Behavioral Goals: Morris [12] listed 17 happiness goals. The left portion of Table II presents them, including goals such as "the inherent happiness that comes with the love of a child," "the competitive happiness of triumphing over your opponents," "the sensual happiness of the hedonist," and so on. Each happiness goal is associated with a type, e.g., the people "the achiever" should have "target happiness," "the winner" should have "competitive happiness," and "the drug-user" should have "chemical happiness."

 ISBN: 978-1-61208-950-8

TABLE II. Happiness goals [12] and their relation to social layers. +'s denote the degree of relevance of each goal to each layer, i.e., Individual, Community, and Social system, respectively. +++: most relevant, ++: moderately relevant, and +: weakly relevant.

				Social Layers		
Category	No.	Name of Happiness	Types	Individual layer	Community layer	Social-system layer
I	8	Painful Happiness	The Masochist	+++		
	11	Tranquil Happiness	The Mediator	+++		
	14	Chemical Happiness	The Drug-taker	+++		
	15	Fantasy Happiness	The Day-dreamer	+++		
II	7	Rhythmic Happiness	The Dancer	+++	+++	
	16	Comic Happiness	The Laugher	+++	+++	
	4	Genetic Happiness	The Relative	+++	+++	
	5	Sensual Happiness	The Hedonist	+++	+++	
III	10	Selective Happiness	The Hysteric	+++	++	
	13	Negative Happiness	The Suffer	+++	++	
IV	9	Dangerous Happiness	The Risk-taker	+++	++	+
	6	Cerebral Happiness	The Intellectual	+++	+++	++
V	1	Target Happiness	The Achiever	+++	+++	+++
	17	Accidental Happiness	The Fortunate	+++	+++	+++
VI	12	Devout Happiness	The Believer		+++	++
	2	Competitive Happiness	The Winner		+++	+++
	3	Cooperative Happiness	The Helper		+++	+++

Kitajima et al. [17] proposed the maximum satisfaction architecture (MSA). MSA assumes that the human brain pursues one of the 17 happiness goals defined by Morris [12] at every moment and switches to another happiness goal when appropriate by evaluating the current circumstances. Each of the happiness goals is associated with one or multiple layers of society. The right portion of Table II presents tentative assignments of the degree of relevance of each happiness goal to each social layer. The middle portion of Figure 1 suggests that any activities for achieving specific behavioral goals would be conducted by individual persons in the pursuit of any of the 17 happiness goals in the social layers presented in the right portion of Table II. Happiness goals define the value structure of the person when he or she makes decisions by running the PCM and memory processes under specific circumstances, while selecting his or her next actions. As such, it is vital to assume the correct happiness goal when supporting the next action selection process of a person, to accomplish the behavioral goals.

There could be associations between the processes of accomplishing behavioral goals and the recognized happiness goals, which could be useful to connect a behavioral goal with a happiness goal. The associations, however, could be vary among individuals. A single observed behavior under a behavioral goal, described in terms of the four operation modes and four processing modes, may have multiple associations with the happiness goals. This is because the condition for feeling happiness is strongly related with individual experiences and the manner in which the reward system functions for that experience [18]. Therefore, the mappings between the behavioral and happiness goals have significant individual and situational differences; a single person could feel different types of happiness when accomplishing a single behavioral goal for different contexts.

2) Guidelines for Supporting Goal Level Interactions: The purpose of designing artifacts has to be linked with any of the SDGs. The design space for artifacts could be explored strategically by associating the targeted SDGs with possible happiness states the user may want to achieve, which is indirectly connected with the behavioral goal of the user [6]. The mode and process level support are truly at the level at which the user could directly interact with. However, the goal-level support is at the level of motivation. The types of happiness goals have discernible aspects of behavioral ecology characterized by individual and contextual differences. Therefore, goal- and contextual-dependent support are needed.

The happiness goals listed in Table II are sorted into six categories according to the degree of relatedness with the social layers, i.e., individual, community, and social-system layers. The categories roughly define the context that the associated behavioral goals are trying to accomplish. The happiness goals in category I could be accomplished individually without any connections with the community or social-system. Those in the category II could be accomplished individually or with the members the individual belongs to. The rest of the categories could be characterized in a similar way. The interface for supporting happiness goals could be designed by category.

Guideline [E]

1. *Provide individually appropriate support for the identified happiness goal that the user might hold when trying to accomplish the behavioral goal.*
2. *Provide contextually appropriate support for the social layer in which the interaction might be conducted.*

 ISBN: 978-1-61208-950-8

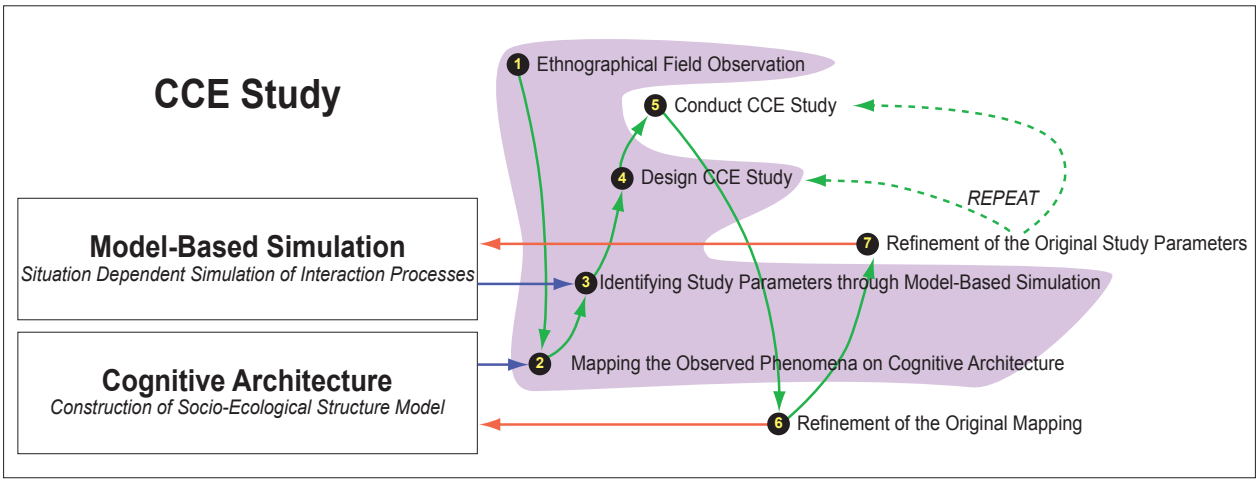


Figure 2. The CCE procedure [2, Figure 5.1].

IV. Application of the Guidelines: A Hint

This section introduces a methodology for conducting field experiments to understand human behaviors as a hint for carrying out a strategically principled search in the design space to obtain design specifications that conform to the guidelines this study proposes. The methodology, cognitive chronoethnography (CCE) [19], should complement the MHP/RT by providing the real data of human behavior for specific situations that should define constraints on the functioning of PCM and memory processes.

A. CCE for Narrowing Down the Design Space

CCE combines three concepts. "Cognitive" declares that CCE deals with interactions between consciousness and unconsciousness in the PCM cycles. "Chrono (-logy)" is about time ranging from ∼100 ms to days, months, and years, and CCE focuses on these time ranges. "Ethnography" indicates that CCE takes ethnographical observations as the concrete study method because in daily life, the Two Minds of people tend to re-use experientially effective behavioral patterns, which are biases and might have individual and contextual differences.

To conduct a CCE study, study participants (elite monitors) are selected. Each defining study field has values. The study question is "what would certain people do in certain ways in certain circumstances (not average behavior)?" Therefore, elite monitors, certain persons, are selected by consulting the parameter space. In this process, it is necessary that the points in the parameter space, which correspond to the elite monitors, are appropriate for analyzing the structure and dynamics of the study field. The methodology is not for human–artifact interaction but for every aspect of the daily life of human beings. Regarding the relationship between CCE and the design space, CCE focuses on understanding the process of interaction between successful artifacts and users and is intended for existing artifacts. Therefore, it is out of scope to predict the kind of interaction that occurs between the user and a non-existent artifact that no one has discussed. The role of CCE is to enable the design space to be narrowed down by a solid understanding of the success stories of existing artifacts, thereby defining the successful areas of new designs. With that in mind, it will be possible to come up with alternatives to successful artifacts. Whether or not new and innovative artifacts are accepted by users is discussed in another guideline paper published by us [5].

B. CCE Procedure for the Human–Artifact Interaction

Figure 2 shows the seven steps to conduct a CCE study [2, Figure 5.1]. The following describes the CCE steps adapted to human–artifact interaction. Necessary additions appear after the general descriptions for the CCE procedures.

(1) *Ethnographical Field Observation:* Use the basic ethnographical investigation method to clarify the outline of the structure of social ecology that underlies the subject to study.

> The subject of study is to understand the manner in which the existing artifacts in question are used successfully by the current users. The ultimate goal of the artifacts is to achieve any of the SDGs through their use by potential users; the current users may be a part of them. The range of users could be widened by appealing appropriately to the right segments of the users. The users could be characterized as an individual, a member of a community or a social-system. Depending on the social layers the users belong to, the happiness goals that could be achieved could vary. The kinds of SGDs that the artifacts with the current or appropriately enhanced specifications could achieve may be widened or corrected. In this step, it is necessary to clarify the outline of the structure of social ecology in terms of the segment of potential users, the social layer they belong to, and the happiness goals they may achieve by referring to Table II.

(2) *Mapping the Observed Phenomena on Cognitive Architecture:* With reference to the behavioral characteristics of people which have been made clear so far and cognitive architectures, consider what kind of characteristic elements of human behavior are involved in the investigation result in (1).

 ISBN: 978-1-61208-950-8

This study proposes the use of MHP/RT as the cognitive architecture for this step. As this study proposes, the human–artifact interaction is characterized at three levels, i.e., the mode, the process, and the goal levels. This is based on the MHP/RT cognitive architecture. Because the artifacts in question realize successful interactions with the users, it is assumed that their design should conform to the guidelines in specific ways. In this step, it is necessary to describe the manner in which they conform to the guidelines, i.e., the mode-level support provided, the process-level support, and the goal-level support.

(3) *Identifying Study Parameters through Model-Based Simulation:* Based on the consideration of (1) and (2), construct an initial simple model with the constituent elements of activated memories, i.e., meme, and the characteristic PCM processing to represent the nature of the ecology of the study space.

In this step, the "what" question answered in (2) is operationalized by turning it into the "how" question. This is answered by constructing an MHP/RT model that could simulate successful users of the artifact in question. The model could run by specifying (a) the likely happiness goals, (b) the possible modes of the assumed interaction, (c) the possible ways of weak synchronization establishment, and (d) the kinds of memes of the simulated user [20][21]. The successful users could be characterized by combining the values assigned to (a) $\sim$ (d), which constitute the study parameters.

(4) *Design a CCE Study:* Based on the simple ecological model, identify a set of typical behavioral characteristics from a variety of people making up the group to be studied. Then formulate screening criteria of elite monitors who represent a certain combination of the behavioral characteristics, and define ecological survey methods for them.

This step follows the standard CCE procedure.

(5) *Conduct CCE Study:* Select elite monitors and conduct an ethnographical field observation. Record the monitors' behavior. The elite monitors are expected to behave as they normally do at the study field. Their behavior is recorded in such a way that the collected data is rich enough to consider the results in terms of the parameter space and as un-intrusively as circumstances allow.

This step follows the standard CCE procedure.

(6) *Refinement of the Original Mapping:* Check the results of (5) against the results of (2) for appropriateness of the mapping. If inappropriate, back to (2) and redo from there.

This step follows the standard CCE procedure.

(7) *Refinement of the Original Study Parameters:* If the result of (5) is unsatisfactory, go back to (4) and re-design and conduct a revised CCE study, otherwise go back to (3) to redo the model-based simulation with a set of refined parameters.

This step follows the standard CCE procedure.

On completion of the CCE cycle, the existing social ecology that characterizes the successful use of the artifact is understood. This understanding is used to widen the range of successful use of the artifact and contribute to determining the direction of strategic development for the maximum utilization of the artifact.

V. CONCLUSION

The purpose of this study was to contribute to realizing a sustainable society of human beings and artifacts. The focus was on the human–artifact interaction, which occurs at the interface between human beings (the interface is composed of multiple autonomous systems, i.e., PCM and memory systems), and artifacts, which are a collection of autonomous systems. This study used a theory-based approach to derive guidelines for application when designing artifacts that should realize a sustainable society.

The constraints imposed on the derivation were: 1) the ultimate purpose of artifacts for realizing a sustainable society should be the achievement of any of the SDGs, and 2) human interactions with the artifacts should be theorized by the MHP/RT cognitive architecture. These constraints were related with each other via the concept of resilience of the interaction processes. On the one hand, the stability of the human–artifact interaction at the mode level, i.e., either System 2 dominant or System 1 dominant processes should be carried out stably, was the necessary condition for the accomplishment of behavioral goals using the artifact. On the other hand, the accomplishment of behavioral goals is linked with the feeling of achieving any of the 17 happiness goals, defined at the three social layers. The behavioral goals do not necessarily have a direct connection with the SDGs; rather, the accomplishment of behavioral goals indirectly contributes to the achievement of any of SDGs as by-products [6]. Because both the happiness goals and SDGs focus on social ecology, the mapping between them could be established [6]. This would complete the links from the stable accomplishment of behavioral goals to the achievement of happiness goals and SDGs for realizing a sustainable society.

Generally, guidelines are useless, unless they are practiced. This study presented a method for applying the derived guidelines based on CCE, which defines the experimentation procedure for complementing the theory of cognitive architecture, MHP/RT. CCE and MHP/RT are the two-wheels of a vehicle to understand the daily behavior of human beings [19]; evidently, the human–artifact interaction is part of it. CCE is used to understand observed behavior. Therefore, it is most useful for extending the existing interaction processes by deliberately extrapolating them by the provision of new interface designs, which should conform to the relevant guidelines. For example, at the process level, weak synchronization has to be realized in the interaction process between the new design and the user. If this interaction is carried out as routine goal-oriented skills in Mode 1, the behavior of the users could be represented as several versions of the GOMS models [9] that are suitable for accomplishing respective behavioral goals. The appropriateness of the new design has to be considered, as to whether it could establish weakly synchronized interaction, given the existing GOMS models.

An artifact is defined as a set of specifications, which are sufficient for engineering to realize a working product. The raison d'être of the artifact would be to contribute to the

 ISBN: 978-1-61208-950-8

achievement of any of the SDGs and to make its users feel any of the happiness goals, to realize a sustainable society through the human–artifact interaction. This study proposed a method for bridging these goals as a set of guidelines on the basis of the scientific understanding of human behavior, provided by the cognitive architecture, MHP/RT, and the methodology for experimentation to narrow down the design space, CCE.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 20H04290. The author would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

[1] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)," Biologically Inspired Cognitive Architectures, vol. 5, 2013, pp. 82–93.

[2] M. Kitajima, Memory and Action Selection in Human-Machine Interaction. Wiley-ISTE, 2016.

[3] D. A. Norman, "Cognitive engineering," in User Centered System Design: New Perspectives on Human-computer Interaction. CRC Press, 1986, ch. 3, pp. 31–61.

[4] H. A. Simon, The Sciences of the Artificial, 3rd ed. Cambridge, MA: The MIT Press, 1996.

[5] M. Kitajima and M. Toyota, "Guidelines for designing artifacts for the dual-process," in Procedia Computer Science, BICA 2015. 6th Annual International Conference on Biologically Inspired Cognitive Architectures, vol. 71, 2015, pp. 62–67.

[6] M. Kitajima, "Cognitive Science Approach to Achieve SDGs," in COGNITIVE 2020 : The Twelfth International Conference on Advanced Cognitive Technologies and Applications, 2020, pp. 55–61.

[7] V. Sarathy, "Real World Problem-Solving," Frontiers in human neuroscience, vol. 12, 2018, p. 261.

[8] A. Newell, Unified Theories of Cognition (The William James Lectures, 1987). Cambridge, MA: Harvard University Press, 1990.

[9] S. K. Card, T. P. Moran, and A. Newell, The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[10] J. L. McClelland and D. E. Rumelhart, Parallel Distributed Processing: Explorations in the Microstructure of Cognition : Psychological and Biological Models. The MIT Press, 6 1986.

[11] D. Kahneman, "A perspective on judgment and choice," American Psychologist, vol. 58, no. 9, 2003, pp. 697–720.

[12] D. Morris, The nature of happiness. London: Little Books Ltd., 2006.

[13] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)," Behaviour & Information Technology, vol. 31, no. 1, 2012, pp. 41–58.

[14] M. Kitajima and M. Toyota, "Four Processing Modes of *in situ* Human Behavior," in Biologically Inspired Cognitive Architectures 2011 - Proceedings of the Second Annual Meeting of the BICA Society, A. V. Samsonovich and K. R. Jóhannsdóttir, Eds. Amsterdam, The Netherlands: IOS Press, 2011, pp. 194–199.

[15] J. Dinet and M. Kitajima, "Immersive interfaces for engagement and learning: Cognitive implications," in Proceedings of the 2018 Virtual Reality International Conference, ser. VRIC '18. New York, NY, USA: ACM, 2018, pp. 18/04:1–18/04:8. [Online]. Available: https://doi.org/10.1145/3234253.3234301

[16] M. Kitajima, J. Dinet, and M. Toyota, "Multimodal Interactions Viewed as Dual Process on Multi-Dimensional Memory Frames under Weak Synchronization," in COGNITIVE 2019 : The Eleventh International Conference on Advanced Cognitive Technologies and Applications, 2019, pp. 44–51.

[17] M. Kitajima, H. Shimada, and M. Toyota, "MSA:Maximum Satisfaction Architecture – a basis for designing intelligent autonomous agents on web 2.0," in Proceedings of the 29th Annual Conference of the Cognitive Science Society, D. S. McNamara and J. G. Trafton, Eds. Austin, TX: Cognitive Science Society, 2007, p. 1790.

[18] M. Kitajima and M. Toyota, "Two Minds and Emotion," in COGNITIVE 2015 : The Seventh International Conference on Advanced Cognitive Technologies and Applications, 2015, pp. 8–16.

[19] M. Kitajima, "Cognitive Chrono-Ethnography (CCE): A Behavioral Study Methodology Underpinned by the Cognitive Architecture, MHP/RT," in Proceedings of the 41st Annual Conference of the Cognitive Science Society. Cognitive Science Society, 2019, pp. 55–56.

[20] M. Toyota, M. Kitajima, and H. Shimada, "Structured Meme Theory: How is informational inheritance maintained?" in Proceedings of the 30th Annual Conference of the Cognitive Science Society, B. C. Love, K. McRae, and V. M. Sloutsky, Eds. Austin, TX: Cognitive Science Society, 2008, p. 2288.

[21] M. Kitajima, M. Toyota, and J. Dinet, "The Role of Resonance in the Development and Propagation of Memes," in COGNITIVE 2021 : The Thirteenth International Conference on Advanced Cognitive Technologies and Applications, 2021, pp. 28–36.

 ISBN: 978-1-61208-950-8

ARTICLE WITHDRAWN BY AUTHORS

 ISBN: 978-1-61208-950-8

 ISBN: 978-1-61208-950-8

Comparison of Visual Attention Networks for Semantic Image Segmentation in Reminiscence Therapy

Liane Meßmer* and Christoph Reich†
Institut for Data Science, Cloud Computing and IT-Security,
Furtwangen University of Applied Science
Furtwangen, Germany
Email: {*l.messmer, †christoph.reich}@hs-furtwangen.de

***Abstract*—Due to the steadily increasing age of the entire population, the number of dementia patients is steadily growing. Reminiscence therapy is an important aspect of dementia care. It is crucial to include this area in digitization as well. Modern Reminiscence sessions consist of digital media content specifically tailored to a patient's biographical needs. To enable an automatic selection of this content, the use of Visual Attention Networks for Semantic Image Segmentation is evaluated in this work. A detailed comparison of various Neural Networks is shown, evaluated by Metric for Evaluation of Translation with Explicit Ordering (METEOR) in addition to Billingual Evaluation Study (BLEU) Score. The most promising Visual Attention Network consists of a Xception Network as Encoder and a Gated Recurrent Unit Network as Decoder.**

Keywords–Visual Attention Networks; Image Caption Generation; Dementia Health Care; BLEU; METEOR.

I. INTRODUCTION

Demand-oriented technical solutions can make a valuable contribution to the care of people suffering from dementia - People With Dementia (PWD)s. Their potential is far from being exhausted. Digital media, which are used today, e.g., on tablets in the context of memory care, have considerable potential for the individualization of care offers, which are also becoming increasingly important because of the increasing differentiation of lifestyles in care for the elderly [1].

Reminiscence therapy is used to address the activation process of people with Dementia [2]. However, the identification of suitable content, as well as the design and evaluation of high-quality reminiscence sessions is very labor-intensive and places high qualification demands on care workers. Suitable individual contents for PWDs currently have to be identified "manually" and evaluated in terms of their suitability. In practice, a very limited pool of standard content is therefore often used. Dynamic response to interaction with residents is also not possible with the tools currently available. The individual activation and care required for high-quality, biography-based care (as opposed to mere occupation) therefore remains a major challenge, despite the extensive availability of digital content today.

Semantic segmentation is a well known technique, in the field of computer vision, and the basis to full understanding of a scene. With the popularity of Deep Learning in recent years, many semantic segmentation problems are being addressed with Deep Learning architectures that far outperform other approaches in terms of accuracy and efficiency. Image description models typically consist of an encoder-decoder architecture. Most commonly, Convolutional Neural Networks (CNN)s are used as encoders for image feature extraction and Recurrent Neural Networks (RNN)s are used as decoders for image description modelling [3]. This work analyses the potential of the Convolutional Neural Networks Inceptionv3, VGG16/19-Net, ResNet101 and Xception in combination with the Recurrent Neural Networks Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), concerning their suitability for use in an image search and selection system for people with dementia.

The activation of PWD requires the selection of pictures according to the following picture characteristics such as color, shapes, amount of objects, meaning according to life themes, etc. Ji et al. [4] present a system that allows images to be grouped together based on a domain-specific ontology. Through this approach, a precise image selection, e.g., using a life topic ontology - can be realized. Jaiswal, Liu and Frommholz [5] describe a selection system with user personalization in which "Information Foraging Theory" is applied to explain the suggestions. This allows users to be made "transparent" as to why an image was selected or suggested in the first place.

The dataset used for the model analysis in the context of PWD must be suitable for the task of semantic image description on the one hand, but also fit the life topic ontology of PWD so that suitable content can be identified to activate them. Due to the wide range of different categories in the Microsoft Common Objects in Context (COCO) dataset, many life themes can be covered, such as animals, people or leisure activities. Furthermore, the dataset provides image descriptions that can be used for feature extraction by a CNN, as well as for semantic image description by the RNN. Therefore, the COCO dataset [6] is particularly suitable in the context of semantic segmentation of image content for PWD.

The aim of the work is to compare different VAT architectures for semantic image segmentation, using CNNs and RNNs in the context of people with dementia, and thereby enable

 ISBN: 978-1-61208-950-8

automatic identification, as well as selection of appropriate activation session content from available digital images. The focus is on image features such as the meaning of life themes, colors, shapes and quantities of objects. In particular, an automated, individual and biography-related media selection improves the quality of the session and relieves the caregivers by shortening the preparation time.

This work consists of 7 sections. Section 2 deals with the related work. Semantic Image Segmentation with Visual Attention Networks (VATs) is described in Section 3. In Section 4, the data used for training is presented and explained. The training process, is described in Section 5. Finally, the results are presented in Section 6 and Section 7 describes the conclusion and future work.

II. Related Work

Alm et al. [7] proposed the first project that developed an Application for digital reminiscence therapy was the Computer Interactive and Conversation Aid (CIRCA) Project. This project aimed to support Dementia patients with digital reminiscence sessions. Over the years, it was supplemented by different new technologies, like a specific interface for the interaction with the System [8] or a touch screen computer to enable an easier interaction with the system [9]. Today, CIRCA is an interactive multimedia application, which supports digital pictures, video and music. The latest publication from Astell, Smith, Potter and Preston-Jones [10] was the work "Computer Interactive Reminiscence and Conversation Aid groups - Delivering cognitive stimulation with technology" which demonstrates the effectiveness of CIRCA for group interventions.

The work "Interactive memories - technology-aided reminiscence therapy for people with dementia" from Klein and Uhlig [11] published an approach to support reminiscence caregivers in their work, so that appropriate images for the sessions can be selected more easily, namely by automatic labeling of available content. They used mixed reality user interfaces to help the user explore media artifacts of their individual biography and spark conversations with caregivers and family . Therefore, the Multimedia content of a therapy session has to be identified manually, that is where our work comes in. We evaluate different architectures of Visual Attention Networks to automatically describe images that match the biographical content of a dementia patient to facilitate and improve the quality of Reminiscence sessions.

The network architecture in this paper is based on the approach proposed by Xu et al. [12], in the work "Show, Attend and Tell". They describe the architecture of a Visual Attention Network that uses CNNs as Encoder an RNNs as decoder, with an additional attention layer inside of the RNN network. With the attention layer the network is able to select the focus on specific parts of an image instad of processing the image as a whole.

A comparison of different Visual Attention Network architectures is presented by Ankit, Subasish, Anuveksh and Vinay [13] in the work "Image Captioning and Comparison of Different Encoders". They compare different CNNs as Encoder configuration like Inceptionv3, VGG16, VGG19 and InceptionResNetV2. For Training they use the Flickr8k dataset. The result is that an Inceptionv3 Network performes best as Encoder. They compare the results with BLEU Score, similar to our work, but instead of Flickr dataset, we use prepared COCO dataset with our image class for training and in addition we train different RNN Networks (LSTM and GRU) as Decoder.

There are several metrics which can be used to evaluate automatically generated image descriptions. Common used Metrics are BLEU and METEOR, each metric has well known benefits and blind spots. We have to measure correlation to human judgements and evaluation of the syntax in the generated sentences. To adress these two challenges Cui, Yang, Veit, Huang and Belongie [14] propose a novel learning based discriminative evaluation metric, that is directly trained to distinguish between human and machine-generated captions. They conclude, that the metric could be an effective complementary to the existing rule-based metrics.

III. Semantic Image Segmentation with Visual Attention Networks

Visual Attention Networks are used to address the Problem of generating image descriptions in the field of full scene understanding. It's not only necessary to predict the objects shown on an image. Furthermore, the model should be able to capture the relationships between different objects on an image to convert them into a natural language from large sets of data [12].

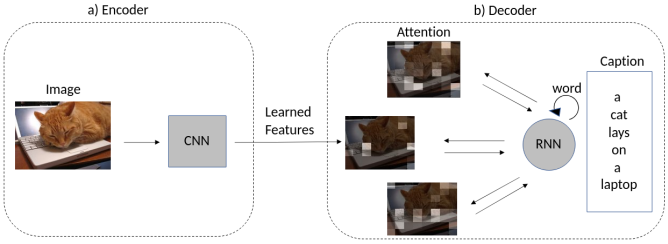


Figure 1. Visual Attention Network Architecture [15]

A VAT consists of an Encoder a) - Decoder b) Architecture as shown in Figure 1. As Encoder a CNN is used to extract the image features into a vectorial representation and a Recurrent Neural Network is used to generate appropriate descriptions for the extracted image features [16][12]. The RNN has a attention layer inside which is based on the functionality of a human visual system. Instead of processing the scene as a whole, the attention is focused on different parts of an image.

IV. Dataset

Reminiscence sessions consist of content that reflects the biographical content of a person with dementia. Media that correspond to the biographical events of a person with dementia can trigger memories which evoke activation's in the patient. These biographical contents are also called life themes and refer to any area of life. Possible examples of life

 ISBN: 978-1-61208-950-8

themes are: Animals, travel, nature, religion, childhood and youth or occupation. The life themes can be represented by different types of media or combinations of them. Examples are pictures, music or videos. The media that correspond to the life themes do not necessarily evoke positive memories in the person with dementia. Therefore, it is important to know negative memories and fears of a patient, so that media content which trigger these emotions can be sorted out for a reminiscence session.

To match the life themes of dementia patients, our dataset contain classes with objects from everyday life, like the classes in MS COCO dataset. The MS COCO dataset [6] has many classes in common with the dementia patients life themes, so this dataset is used for training. For image captioning, each image from the COCO dataset is described with 5 different sentences.

This work primarily targets the description of dog and cat images, so these categories are filtered from the dataset. Dog images have two categories in the area of Reminiscence Therapy: Dog images that activate positive memories in the PWD and dog images that might trigger negative memories. For example dangerous looking dogs, aggressive dogs or snarling dogs are objects that can evoke negative memories. In total, 4114 images of the category "cat" and 4385 images of the category "dog", with a total of 42495 image descriptions, are filtered from the dataset. COCO only contains images with friendly looking dogs but we need angry, fear producing dogs, too. So, the dataset is extended with the category "angry dogs" and filled with new image content, from free image databases. Similar to the caption style of the MS COCO dataset, we labeled each image with 5 description sentences. Every sentence is natural language formulated and contains one or more of the following information: A dog shows or bares his teeth, Number of dogs in the picture, Color of the dogs, Background color, Meadow in background or other objects and toys on the picture.

The number of images in this category amounts to 360 training images with 1800 descriptions. In total, our dataset consists of 8859 images, with 44295 image descriptions. For training, we use a random $80/20$ split on the dataset, to split it into train and validation set.

V. Evaluation and Comparison

This section describes the technologies and parameters used for training. Furthermore, the results are presented, evaluated and compared.

A. Training

For training of the networks, we use the described dataset. The networks are all used with weights pre-trained on ImageNet dataset. We use a fixed length image caption of 9 words per sentence, because performance is poor on long input or output sequences. The Training Vocabulary consists of all words that occur more than tree times in the vocabulary, In total there are 6660 words in the training Vocabulary. Unknown words are provided with the token $< unk >$. The Networks are trained with 100 Epochs.

B. Encoder and Decoder

The use of a CNN as Encoder in a Visual Attention Network Architecture was successfully evaluated by recent works [17][18]. We evaluate different CNN implementations for image feature extraction to determine which is the best in the area of image caption generation for reminiscence sessions. The networks compared and evaluated are: Inceptionv3 [19], ResNet101 [20], VGG16/19 [21] and Xception [22] as Encoder. As Decoder LSTM [23] and GRU [24] networks are used.

C. Metrics for image captioning evaluation

For a formal comparison of the captions generated by the VATs, different metrics are used.

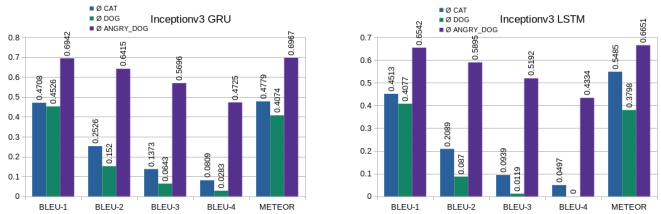
BLEU Score is a metric that can be used to automatically evaluate machine-generated image captures. BLEU is fast, inexpensive and language independent. The metric correlates strongly with the reference captions, as the caption length, word choice and word order are used to calculate the BLEU Score [25].

METEOR is a metric for formal and automatic evaluation of machine-generated captions, too. The metric measures not only precision (accuracy of the match), but also recall (completeness of the match), unlike the BLEU Metric. In this metric, word agreement is not determined using n-grams, but using unigrams, which are grouped into as few chunks as possible, where a chunk is defined as a set of unigrams that are adjacent in the hypothesis and the reference. The metric solve some problems of the BLEU Metric. BLEU measures correlation at the corpus level and METEOR also measures correlation with human judgment at the sentence or segment level [26].

D. Results

The results calculated by the metrics are represented in the following figures. They show respectively the results of the BLEU Score with different n-grams and the results of the METEOR metric. Figure $a)$ represents the results obtained by using the GRU implementation as decoder and $b)$ shows the results of LSTM network as decoder. For evaluation, a dataset, containing 10 images for each class (cat, dog, angry dog), was created. Each image is described with 5 captions per image as reference to calculate the metric scores.

Figure 2 shows the results of Inceptionv3 encoder.



(a) Results Inceptionv3 GRU

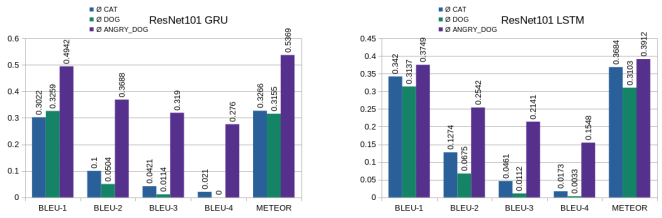
(b) Results Inceptionv3 LSTM

Figure 2. Results Inceptionv3

The results are a little better when using a GRU decoder as opposed to the results of the LSTM decoder, regardless of the metric. The class angry_dog performs best, while BLEU with the use of a 4-gram gives the worst results. However, depending on the class, the results are equally distributed for both the GRU and the LSTM network.

Figure 3 shows the results of the ResNet101 Encoder. This Encoder produces the worst results in comparison to all other Encoder - Decoder combinations. Whereby the GRU decoder provides better results, except for the dog class, calculated with the BLEU score using a 1-gram.

Class dog and angry_dog have nearly the same METEOR value with LSTM Decoder. This phenomenon does not occur with any other network.



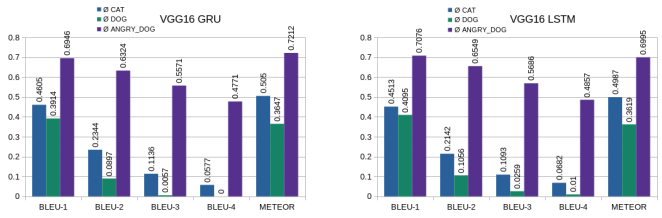
(a) Results ResNet101 GRU (b) Results ResNet101 LSTM

Figure 3. Results ResNet101

The evaluation results of the trained VGG16 Network are shown in Figure 4. With this network architecture, the results are still stable with different decoders. The values of the metrics hardly differ when using the GRU decoder compared to the use of the LSTM encoder.

The results of VGG16 are generally better than those of Inceptionv3 and ResNet101. But in contrast to VGG19 and the Xception rather worse, except for the angry_dog class, trained with LSTM decoder.

This class performs best among all network architectures. There is also an exception in the cat class, which was calculated with the BLEU score and a 4-gram. In this class, the network also performs better than all others used for this evaluation.



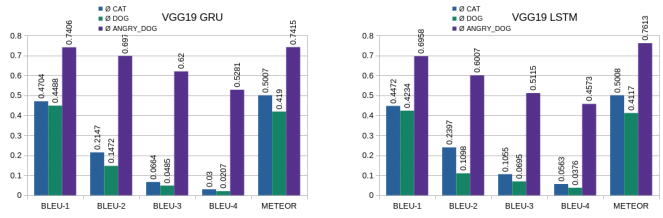
(a) Results VGG16 GRU (b) Results VGG16 LSTM

Figure 4. Results VGG16

Figure 5 shows the results of VGG19 Encoder. For our use case, the deeper VGG network with 19 layers performed better than the VGG network with 16 layers. VGG16 performs better for only one class (which may be due to the natural variance of an RNN network), VGG19 performs better in multiple classes and by using GRU and LSTM decoders.Co

In contrast to all other architectures, BLEU-4 and METEOR give the best results in the angry_dog class, using a GRU decoder. By using a LSTM decoder the values vary and the class Dog calculated by BLEU with a 4-gram performs best, as well as the class angry_dog by calculating the METEOR value.

Both VGG networks have the property that they provide stable results no matter which decoder is used, the results do not vary much. Furthermore, the result values are equally distributed, with all data classes.



(a) Results VGG19 GRU (b) Results VGG19 LSTM

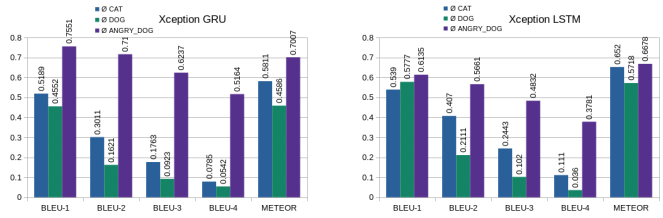
Figure 5. Results VGG19

The last network is the Xception Network. The results are shown in Figure 6. The values of this network are the best, no matter which decoder is used. Both GRU and LSTM results are better than the values from the other networks.

The only exception is within the angry_dog class and LSTM decoder. It does not outperform the VGG16 network by using BLEU score and METEOR metric for result evaluation.

The Xception network is based on the architecture of the Inception network. However, the inception modules are replaced by depth wise separable convolutions followed by point wise convolutions. The number of parameters is the same in both networks, but the Xception network uses the parameters in a more efficient way than the Inception network [22].

As our results show, the extended Inception architecture is successful, since the results are better than those of the Inception network.



(a) Results Xception GRU (b) Results Xception LSTM

Figure 6. Results Xception

For all networks the caption generation for the class ”angry_dog” works best. This is because we created and labeled

the class ourselves. So, the labels used for training fit our use case better than the labels from COCO dataset.

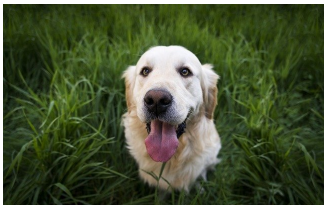
Generally, the results get worse with increasing n-gram value and the METEOR results are quite uniform due to the use of the chunks.

E. Human Evaluation

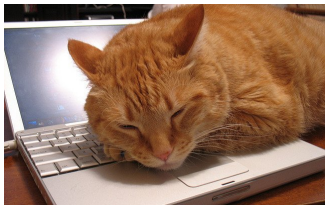
Complementing the evaluation using metrics, we conducted a human evaluation of the results to determine whether the results are good enough to make the trained weights applicable in practice for people with dementia.

In humans, the formal translations exist only in their mind. So, people's pattern translations are preverbal representations, and can be realized with several synonymous expressions when they are verbalized. Therefore, human evaluators may equally evaluate different translation variants as "correct," although their evaluations might differ depending on their emphases [27]. That means, even results with a low BLEU or METEOR value could be rich picture descriptions, since many words and sentence positions mean similar things, even if they are formally different from each other.

We picked one image from each of the ten images that are in a test class and compared the result captures. For this purpose, we collected the best and the worst results generated by the network. The best results are shown in Table I and the worst in Table II. The results refer to Figure 7, where a), b) and c) are respectively selected images from the test classes that were present in the test dataset. The results must refer to the objective opinion of a person and not to a technical evaluation.



(a) Evaluation dog picture



(b) Evaluation cat picture

(c) Evaluation angry dog picture

Figure 7. Example evaluation pictures

The best results based on human evaluation were obtained by the VGG19 and Xception networks. The objects contained in the images were well recognized and correctly described by the networks. Also syntactic criteria are mostly fulfilled by the networks in a sufficient quality. It was observed that the results of the Xception network are more "stable" than the results of the VGG19 network, regardless of which decoder is used. More stable means that the number of the same caption is higher than the number of the same caption generated by other networks, because captions of a RNN can vary, since such a model does not have a fixed number of hidden layers.

TABLE I. HUMAN EVALUATION BEST RESULTS

Picture	Encoder	Caption
a)	Inceptionv3	a white dog wearing a green and holds a
	ResNet101	a dog has a red collar is standing near
	VGG16	a black and white dog
	VGG19	a dog laying in a grassy field looking grass
	Xception	a dog is sitting
b)	Inceptionv3	a cat is curled up asleep lying on a
	ResNet101	a cat is using a laptop keyboard
	VGG16	cat laying on a laptop computer
	VGG19	a orange cat sits on a laptop
	Xception	a cat is laying down on a laptop
c)	Inceptionv3	a black and brown dog looks angry while baring
	ResNet101	a black and brown dog baring his teeth on
	VGG16	an angry looking black and brown dog shows his
	VGG19	an angry looking black and brown dog shows his
	Xception	An angry black and brown dog baring his teeth

TABLE II. HUMAN EVALUATION WORST RESULTS

Picture	Encoder	Caption
a)	Inceptionv3	a cute hair that its mouth resting while wearing
	ResNet101	a dog <unk> in a pink flower and green
	VGG16	a big panting dog
	VGG19	the dog is looking to above his mouth
	Xception	a golden puppy leash standing near some frisbee
b)	Inceptionv3	an orange cat sits resting its head on the
	ResNet101	a cat sleeping half lake in an open suitcase
	VGG16	an orange cat resting it's camera
	VGG19	a cat sleeping half on
	Xception	a close up of a cat sleeping half on
c)	Inceptionv3	a black and brown dog shows his teeth on
	ResNet101	a cat is greeting each other in a chair
	VGG16	a brown dog baring his teeth on green grass
	VGG19	a black and brown dog looks angry while baring
	Xception	an angry looking black and brown dog shows his

The worst results are obtained by ResNet101, the results vary strongly among each other and the network generates many "outlier" captions, which do not fit in any way to the image content that should be described. For example, the caption "a dog has a hat on the beach" or "a small cat is sitting on the ground" are outlier result generated by caption generation with respect to Figure 7 b).

VI. Conclusion and Future Work

This work takes up the basic functionality of Visual Attention Networks and presents their use based on different network configurations intending to automatically describing images in such a way that they can be assigned to the life themes of dementia patients. In this way, reminiscence sessions can be automatically created with biography-related content. By automatically compiling sessions, caregivers are relieved and can invest more time with the patients than in creating the reminiscence session content.

 ISBN: 978-1-61208-950-8

We have figured out which network architecture performs best. For comparison we used the encoder networks Inceptionv3, ResNet101, VGG16/19 and Xception (image feature extraction) in combination with the decoder networks LSTM and GRU (caption generation). For training, we created a dataset specifically suited to dementia patients, which is composed of some classes from COCO dataset and a separate class. By evaluating the networks trained using our dataset, we found that the Xception network in combination with a GRU network produced the best results. Both are evaluated formally and by human.

In the future, the system can be extended with other digital media types, like music or videos. The dataset we use only covers the life theme "animals". To make the reminiscence sessions more valuable, other life themes should be included by extending the training dataset. From a technical point of view, the VAT could be further adjusted by hyperparameter tuning to improve the results and to reduce the number of outlier captions.

REFERENCES

[1] F. Meiland, A. Innes, G. Mountain, L. Robinson, H. van der Roest, A. García-Casal, D. Gove, J. R. Thyrian, S. Evans, R.-M. Dröes, F. Kelly, A. Kurz, D. Casey, D. Szczesniak, T. Dening, M. Craven, M. Span, H. Felzmann, M. Tsolaki, and M. Franco, "Technologies to support community-dwelling persons with dementia: A position paper on issues regarding development, usability, effectiveness and cost-effectiveness, deployment, and ethics," JMIR Rehabil Assist Technol, vol. 4, 01 2017, p. e1.

[2] A. A. Khait and J. Shellman, "Uses of Reminiscence in Dementia Care," Innovation in Aging, vol. 4, no. Supplement_1, 12 2020, pp. 287–287.

[3] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. A. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019, pp. 11 135–11 145.

[4] Z. Ji, W. Yao, H. Pi, W. Lu, J. He, and H. Wang, "A survey of personalised image retrieval and recommendation," in Theoretical Computer Science - 35th National Conference, NCTCS 2017, Wuhan, China, October 14-15, 2017, Proceedings, ser. Communications in Computer and Information Science, vol. 768. Springer, 2017, pp. 233–247.

[5] A. K. Jaiswal, H. Liu, and I. Frommholz, "Effects of foraging in personalized content-based image recommendation," CoRR, vol. abs/1907.00483, 2019.

[6] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," CoRR, vol. abs/1405.0312, 2014.

[7] N. Alm, A. Astell, M. Ellis, R. Dye, G. Gowans, and J. Campbell, "A cognitive prosthesis and communication support for people with dementia," Neuropsychological Rehabilitation, vol. 14, no. 1-2, 2004, pp. 117–134.

[8] G. Gowans, R. Dye, N. Alm, P. Vaughan, A. Astell, and M. Ellis, "Designing the interface between dementia patients, caregivers and computer-based intervention," The Design Journal, vol. 10, Mar 2007, pp. 12–23.

[9] A. Astell, M. Ellis, L. Bernardi, N. Alm, R. Dye, G. Gowans, and J. Campbell, "Using a touch screen computer to support relationships between people with dementia and caregivers," Interacting with Computers, vol. 22, 07 2010, pp. 267–275.

[10] A. Astell, S. Smith, S. Potter, and E. Preston-Jones, "Computer interactive reminiscence and conversation aid groups—delivering cognitive stimulation with technology," Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 4, no. 1, 2018, pp. 481–487.

[11] P. Klein and M. Uhlig, "Interactive memories: Technology-aided reminiscence therapy for people with dementia," in Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, ser. PETRA '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1–2.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proceedings of the 32nd International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057.

[13] A. Pal, S. Kar, A. Taneja, and V. Jadoun, "Image captioning and comparison of different encoders," Journal of Physics: Conference Series, vol. 1478, 04 2020, p. 012004.

[14] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," pp. 5804–5812, 2018.

[15] L.-M. Meßmer and C. Reich, "Potentials of semantic image segmentation using visual attention networks for people with dementia," pp. 234–252, 2021.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," CoRR, vol. abs/1411.4555, 06 2014, pp. 3156–3164.

[17] J. Aneja, A. Deshpande, and A. Schwing, "Convolutional image captioning," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 5561–5570.

[18] S. Katiyar and S. K. Borgohain, "Analysis of convolutional decoder for image caption generation," CoRR, vol. abs/2103.04914, 2021.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," pp. 2818–2826, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 1800–1807.

[23] M. Phi, "Illustrated Guide to Recurrent Neural Networks - Towards Data Science," Medium, Jun 2020, [Retrieved: Jan, 2022]. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9

[24] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, Mar 2020, p. 132306.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, jul 2002, pp. 311–318.

[26] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, jun 2005, pp. 65–72.

[27] H.-Y. Chung, "Automatische Evaluation der Humanübersetzung: BLEU vs. METEOR," Lebende Sprachen, vol. 65, no. 1, Apr 2020, pp. 181–205.

 ISBN: 978-1-61208-950-8