CL-XAI: Toward enriched Cognitive Learning with eXplainable Artificial Intelligence

Muhammad Suffian[1][0000−0002−1946−285X] Ulrike Kuhl[2][0000−0002−9405−918X]
Jose M. Alonso-Moral[3][0000−0003−3673−421X] and Alessandro
Bogliolo[1][0000−0001−6666−3315]

[1] Department of Pure and Applied Sciences, University of Urbino, Urbino, Italy
`m.suffian@campus.uniurb.it, alessandro.bogliolo@uniurb.it`
[2] Research Institute for Cognition and Robotics, Bielefeld University, Bielefeld,
Germany
`ukuhl@techfak.uni-bielefeld.de`
[3] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
`josemaria.alonso.moral@usc.es`

Abstract. As computational systems supported by artificial intelligence (AI) techniques continue to play an increasingly pivotal role in making high-stakes recommendations and decisions across various domains, the demand for explainable AI (XAI) has grown significantly, extending its impact into cognitive learning research. Providing explanations for novel concepts is recognised as a fundamental aid in the learning process, particularly when addressing challenges stemming from knowledge deficiencies and skill application. Addressing these difficulties involves timely explanations and guidance throughout the learning process, prompting the interest of AI experts in developing explainer models. In this paper, we introduce an intelligent system (CL-XAI) for Cognitive Learning which is supported by XAI, focusing on two key research objectives: exploring how human learners comprehend the internal mechanisms of AI models using XAI tools and evaluating the effectiveness of such tools through human feedback. The use of CL-XAI is illustrated with a game-inspired virtual use case where learners tackle combinatorial problems to enhance problem-solving skills and deepen their understanding of complex concepts, highlighting the potential for transformative advances in cognitive learning and co-learning.

Keywords: Cognitive Learning · Explainable AI, · Human-centered AI · Problem Solving · Counterfactual Explanations · Co-Learning.

1 Introduction

In the realms of learning theory, artificial intelligence (AI), and human-computer interaction (HCI), the pursuit of problem-solving and optimal solution finding have historically been perceived through distinct lenses for machines and humans

[12, 28, 32]. Machines, equipped with their computational capabilities and cost-driven optimisation, operate in a space detached from human cognition [17]. Conversely, humans rely on their unique problem-solving approaches, drawing from experiences and intuition [16]. This separation contradicts the principles of co-learning and effective HCI. In our contemporary era, where AI systems often outperform humans in various domains, bridging this gap is imperative to create a more enriching learning environment [7]. The underexplored domain of human-machine co-learning, aiming to foster mutual improvement and progress, must be addressed. Humans need an opportunity to plunge into the intricate inner workings of AI's cognitive machinery, actively participating in co-learning to solve complex problems. In turn, AI systems should reap the benefits of human wisdom, leveraging human input and feedback to alleviate the computational burdens that once restrained their problem-solving prowess.

As AI systems play an increasingly pivotal role in making high-stakes recommendations and decisions across various domains, the demand for eXplainable AI (XAI) to elucidate the rationale behind these systems grows [1]. This paper introduces the so-called Cognitive Learning with eXplainable AI (CL-XAI) system. We explore the potential of co-learning with counterfactual explanations (CEs), where humans and machines collaborate in problem-solving tasks. CEs enable users to grasp the "what if" aspect of AI decisions, shedding light on alternative courses of action and improving transparent communication [30].

In this context, our focus is twofold: firstly, we investigate how human learners' cognitive learning processes are influenced when they use the CL-XAI tool to receive explanations. Secondly, we rigorously assess the effectiveness of the CL-XAI tool by gathering feedback from humans to evaluate its accuracy and helpfulness in providing explanations. The co-learning experience unfolds with regular interactions between learners and an explanation tool. This tool bridges human intuition and machine logic, offering learners a lifeline in their pursuit of optimal solutions. If a human learner successfully identifies an optimal solution, then it is an evidence to the acquisition of knowledge comparable to that acquired by a machine learning (ML) model [19]. For the co-learning experience, we propose a virtual game-inspired scenario where learners solve combinatorial problems to achieve improved results (see section 3). The learner receives explanations at regular intervals to solve the task, and the log of the learner's choices and attempts is recorded to evaluate its mental model. Overall, we believe that the synergy between human cognition and XAI guidance holds the promise of transformative advances in cognitive learning, a step towards co-learning, ultimately bolstering problem-solving skills and fostering a comprehensive understanding of complex concepts.

The rest of the paper is structured as follows. Section 2 provides the necessary background. Section 3 outlines the CL-XAI system, elucidating our approach to incorporating XAI into the cognitive learning process. Section 4 delves into subjective evaluation measures for the proposed framework, and finally, in section 5, we draw conclusions from our work, emphasising potential applications and the avenues for future research in the realm of XAI-driven cognitive learning.

2   Background

Cognitive learning is a pedagogical approach that emphasises the development of comprehensive mental models among learners [10]. Such mental models play a pivotal role in knowledge acquisition and problem-solving, enabling individuals to navigate complex domains effectively [26]. The collaboration between humans and machines in cognitive learning, often called co-learning, has great potential [6, 13]. It has been shown to enhance problem-solving skills and deepen the understanding of complex concepts [13]. On the one hand, prior studies have introduced theoretical frameworks for symbiotic learning systems [31]. These frameworks depict a reciprocal learning process, where the learner acquires knowledge from the system, and, conversely, the system gains insights from the learner, facilitated through reinforcement learning. On the other hand, research into explainable recommendation systems has expanded to education. For instance, Barria-Pineda et al. [3] have explored the domain of recommending resources in programming classes. Tsiakas et al. [25] have investigated using cognitive training recommendations for primary and secondary school children.

In addition, in the field of mathematics [18] and across diverse scientific domains, including chemistry [5], educators routinely employ worked examples. These pedagogical tools serve a dual purpose: they provide solutions and offer explanations rooted in an expert's mental framework, which can be comprehended from a novice's standpoint [23, 24]. Conversely, our approach integrates an XAI tool for generating explanations, which are subsequently provided to novice learners to facilitate the construction of their mental models. This method constitutes an automated system that circumvents the necessity for an expert's mental model, as it leverages the pre-existing capabilities of the XAI system.

3   CL-XAI

CL-XAI is a tool that encompasses three different components: (1) The explanation method called User Feedback-based Counterfactual Explanations (UFCE) [22] to assist learners in solving a given combinatorial problem. (2) The virtual *Alien Zoo* framework [11] is utilised to design the task for the learners as a use case. (3) A web-based game-inspired user study is designed for learners to enhance their learning and knowledge about the artefacts of the underlying AI model when solving the given combinatorial problem.

3.1   UFCE

The field of XAI encompasses various technical approaches aimed at enhancing the transparency, interpretability and explainability of AI systems [2]. One prominent method involves using local post-hoc interpretability approaches for elaborating explanations of black-box models. Such approaches are usually supported by inherent interpretable white-box models such as linear models, decision trees or rule-based systems. Indeed, providing users with explicit rule

sets governing AI behaviour is a way of ensuring a clearer understanding of the decision-making process. Additionally, techniques such as feature importance scoring, attention mechanisms, and saliency maps enable the identification of crucial model elements and their contributions to the overall outcome. CEs provide insights into how alterations in input data affect the model's outputs, while visual aids like heatmaps and activation maps highlight influential areas in images. Moreover, human-in-the-loop interactions and natural language explanations also contribute to the holistic understanding of AI operations, catering to various interpretation needs and promoting trust in AI decision-making.

In this context, UFCE[4] provides CEs[5] to the learner to devise different strategies to solve the task at hand. UFCE, herein referred to as the "XAI tool", is not self-driven; rather, it is a human-in-the-loop explainer that exploits learner feedback to generate customised explanations. In addition, it guides the learner towards correct input options whenever the given feedback does not produce any explanation. The task is designed by keeping the needs of both the learner and the explainer.

3.2   Use Case and Experimental Context

The efficacy of an explanation is influenced by its purpose and the specific audience it is intended for [15]. These two factors are pivotal in selecting a suitable use case and shaping the experimental environment. Thus, we have adopted the Alien Zoo [11], a web-based game-inspired virtual space wherein the learners do not have prior knowledge and only gain knowledge through the explanations provided by the explainer.

The learner's task is to nurture an Alien so-called 'Shub' in the Alien Zoo by feeding various combinations of plants. In our case, in contrast to the original implementation [11], the well-being of Shub is directly influenced by the choices learners make in selecting plant combinations. The leaves of plants are attached with a specific time cost to find them. The learner is exposed to a combinatorial problem in which the random number of leaves of each plant are given, which are not a better diet for the health of Shub. The learner has to solve this problem by making different combinations of the plants to constitute a nourishing diet adhering to their time cost. The learner-selected plants will be fed to an AI system, which will predict whether this combination can improve Shub's health. The choices made when feeding Shub have immediate outcomes, resulting in poor or better health. The XAI tool guides the learner in making optimal decisions. At regular intervals, learners are provided with explanations alongside their previous selections. These explanations highlight a choice that could have yielded a more favourable outcome. Also, these explanations are, by default, enriched with hints about the underlying data distribution associated to the prediction model and about how Shub's health improves, ultimately assisting learners in making informed decisions and enhancing their mental model.

---

[4] https://github.com/msnizami/UFCE
[5] In the rest of the paper, we use explanations and counterfactual explanations (CEs) interchangeably.

3.3 Game-inspired User Study

A game-inspired user study is designed to record the user solutions for the given tasks and explanations provided at regular intervals. The data is collected through game locks (logs) for the analysis and evaluation of the cognitive learning process of learners and the extent of goodness of provided explanations by the XAI tool. The user interface (UI) is meticulously designed to offer learners an intuitive and engaging experience throughout their journey of nurturing Shub and exploring the intricate relationships between plant combinations and growth outcomes. The UI comprises various screens, each carefully crafted to provide essential information and interactivity. In the right hand side of Fig. 1, an avatar represents Shub and dynamically reflects its fitness level on a vertical bar as the chosen plant combination impacts Shub's health. The bar limits indicate the optimal and unsatisfactory fitness levels. On the left, the five different plants are displayed on top, which constitute the diet, and a test input is shown at the bottom, which needs the learner's attention to customise it by selecting any combination from the drop-down menu's given for each plant. The user can also see the available time to improve the health with the number of rounds and the time required to invest in one leaf of each plant.
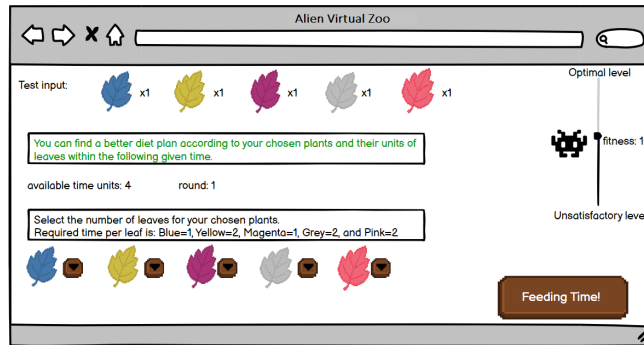
Fig. 1. The learner's task is to go for those combinations of plants by selecting from drop-down menus that can be searched in the available time. After each search/selection, the learner feeds the plants to Shub and waits for the outcome (which is displayed in the next screen, see Fig. 2).

Technical Details. The realisation of the Alien Zoo entails a rigorous segregation between the front end, responsible for crafting the game interface that participants interact with, and the back end, delivering the predictions made by the AI system along with the explanations provided by the XAI tool. The web interface utilises Phaser3[6], an HTML5 game framework driven by JavaScript. The system's backend, on the other hand, is founded on Python3, leveraging

---

[6] https://phaser.io/
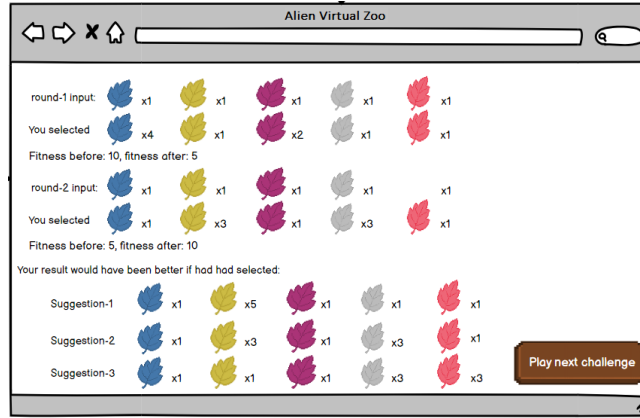
Fig. 2. In this screen, the learner can see the selection of plants throughout the different rounds and their outcome for fitness (better or worse). Additionally, the learner receives different suggestions (explanations). If the learner had used these suggestions, this would have been a better diet for Shub.

the sklearn[7] package for supporting ML algorithms. The underlying ML model is trained using synthetic plant data [11], and it predicts the fitness of Shub, thereby influencing the game's dynamics. The learner input reaches this model via the front end, allowing an analysis of the potential for yielding positive outcomes and consequently enhancing Shub's fitness. This Python-based framework is adept at generating CEs with UFCE to ensure adaptability and accommodate various ML algorithms.

4   Subjective Evaluation Measures

In evaluating XAI systems, examining specific cognitive states or processes, herein referred to as "cognitive metrics", is a central concern [9]. These metrics are instrumental in assessing whether learners have achieved a pragmatic understanding of the AI system, particularly in light of the explanations furnished by the XAI tool. Drawing from established approaches in cognitive science and psychology [9], Hoffman et al. [8] proposed a conceptual model elucidating the explanation processes of an XAI system and how a learner's pragmatic comprehension of these explanations can be assessed across distinct functional stages. Hoffman's model delineates three pivotal functional stages within the XAI system's operations: *explanation generation*, *learner's mental model generation*, and *learner's enhanced performance resulting from the assimilated mental model*. Consequently, the evaluation of XAI systems through cognitive metrics can be systematically framed within these three stages, offering a comprehensive approach to gauge the effectiveness of these systems in facilitating user understanding and performance improvement.

---

[7] https://scikit-learn.org/stable/

In accordance with Hoffman et al. [8], during the explanation generation phase, an assessment of the learner's practical comprehension of the AI system can be made by examining the learner's cognitive processes, which gauge the quality of the explanation (referred to as "Explanation Goodness") and the degree of satisfaction with it (referred to as "User Satisfaction"). Between the explanation generation phase and the subsequent stage of constructing the learner's mental model, which is influenced by the explanations received, learners gradually update their mental models, with several psychological factors potentially influencing the model-building process. Evaluating the extent of a learner's understanding and satisfaction poses a formidable challenge. To address this challenge, we have extended the Alien Zoo, where the learners are now assigned problem-solving tasks, and their performance in solving these tasks serves as a metric to determine the level of understanding of the provided explanations. This, in turn, corroborates the effectiveness of the XAI tool in generating high-quality explanations. User Satisfaction can be assessed using subjective measures through the administration of a questionnaire, with for example the *Explanation Satisfaction Scale* introduced by Hoffman et al. [8] and later refined by van der Waa et al. [29]. This scale provides a reliable and psychometrically robust means of gauging user satisfaction with a system's explanations.

User understanding, in the context of XAI, pertains to the development of a learner's "mental model" of a system's inner workings [9]. The concept of a "mental model" draws from psychological theories, denoting an individual's internal representation of the people, objects, and environments with which they interact [20, 21]. In the realm of XAI, for our case, an ideal outcome is for the learner's mental model to reflect the XAI tool accurately. Explanations play a pivotal role in facilitating the construction of precise mental models, which can be categorized as follows: global understanding, signifying a general comprehension of a system's functioning; local understanding, signifying insight into a specific decision made by the system; and functional understanding, representing a grasp of the system's capabilities and intended uses [8, 9]. In this work, our goal is to capture the local understanding of the learner's mental model.

In cognitive psychology research, learners' performance in task-learning endeavours is subject to the influence of many variables. These encompass factors such as age [4], learning experience [28], cognitive abilities [10], cultural and socio-emotional factors [27]. Factors related to mental health [14] collectively constitute critical determinants.

Proof of Concept In our ongoing user study, the focus is on cognitive learning and co-learning mechanisms. We aim to investigate 4 key measures:

– **Explanation Goodness**. We evaluate the quality of explanations provided to learners, aiming to assess how well they convey complex data relationships. We anticipate that well-crafted explanations will positively influence other measures.

– **User Satisfaction**. Post-game surveys collect learner feedback on their satisfaction with the explanations. Learners' assessments of the explanations'

comprehensibility, usefulness, and overall satisfaction provide insights into user satisfaction.

– **User Understanding**. Our primary inquiry is whether users can enhance their understanding of complex data relationships through explanations. By providing CEs, we aim to help users better grasp the system's intricacies.

– **Task Learning**. We expect that improved user understanding, facilitated by the explanations, will lead to task learning. Task learning is assessed through metrics such as fitness levels during the game and the time taken to make decisions, reflecting learners' increased ability to identify crucial data factors and make informed choices.

By examining these measures, we aim to uncover how explanation quality influences cognitive learning and co-learning mechanisms with the CL-XAI tool.

5   Conclusion

In this paper, we have introduced the CL-XAI tool, designed to facilitate cognitive learning with XAI. The outcomes of our research will offer significant potential for enhancing the development and refinement of XAI techniques, ultimately leading to improved cognitive learning experiences.

Our deliberate focus on the convergence of XAI and cognitive learning stems from the recognition that learners, especially in educational and training contexts, stand to gain substantially from AI systems that are understandable and transparent. Our work, centred on providing timely and lucid explanations for intricate concepts and problem-solving tasks (combinatorial problems), seeks to empower learners, bridge knowledge disparities, and cultivate a deeper understanding of challenging subject matter. Furthermore, the implications of this research extend beyond the realm of education to encompass domains where human-AI collaboration is pivotal, such as healthcare diagnostics, legal decision support, and financial analysis.

In summary, we assert that the synergy between human cognition and the guidance offered by the CL-XAI tool holds the potential for transformative advancements in cognitive learning. This will mark a significant stride toward co-learning, ultimately fortifying problem-solving abilities and nurturing a comprehensive grasp of complex concepts.

Acknowledgments

Bibliography

[1] Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access **6**, 52138–52160 (2018)

[2] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Ser, J.D., Díaz-Rodríguez, N., Herrera, F.: Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion p. 101805 (2023). https://doi.org/10.1016/j.inffus.2023.101805

[3] Barria-Pineda, J., Akhuseyinoglu, K., Želem-Ćelap, S., Brusilovsky, P., Milicevic, A.K., Ivanovic, M.: Explainable recommendations in a personalized programming practice system. In: International conference on artificial intelligence in education. pp. 64–76. Springer (2021)

[4] Chan, C.Y.H., Chan, A.B., Lee, T.M.C., Hsiao, J.H.: Eye-movement patterns in face recognition are associated with cognitive decline in older adults. Psychonomic bulletin & review **25**, 2200–2207 (2018)

[5] Crippen, K.J., Earl, B.L.: The impact of web-based worked examples and self-explanation on performance, problem solving, and self-efficacy. Computers & Education **49**(3), 809–821 (2007)

[6] Deiss, O., Biswal, S., Jin, J., Sun, H., Westover, M.B., Sun, J.: HAMLET: Interpretable human and machine co-learning technique. arXiv preprint arXiv:1803.09702 (2018)

[7] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: When will AI exceed human performance? evidence from AI experts. Journal of Artificial Intelligence Research **62**, 729–754 (2018)

[8] Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)

[9] Hsiao, J.H., Ngai, H.H.T., Qiu, L., Yang, Y., Cao, C.C.: Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). arXiv preprint arXiv:2108.01737 (2021)

[10] Johnson-Laird, P.N.: Mental models and cognitive change. Journal of Cognitive Psychology **25**(2), 131–138 (2013)

[11] Kuhl, U., Artelt, A., Hammer, B.: Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. Frontiers in Computer Science **5**, 20 (2023)

[12] Langley, P.: Intelligent behavior in humans and machines. Advances in Cognitive Systems **2**, 3–12 (2007)

[13] Lieto, A., Radicioni, D.P.: From human to artificial cognition and back: New perspectives on cognitively inspired AI systems. Cognitive Systems Research **39**, 1–3 (2016)

[14] Marin, M.F., Lord, C., Andrews, J., Juster, R.P., Sindi, S., Arsenault-Lapierre, G., Fiocco, A.J., Lupien, S.J.: Chronic stress, cognitive functioning and mental health. Neurobiology of learning and memory **96**(4), 583–595 (2011)

[15] Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS) **11**(3-4), 1–45 (2021)

[16] Newell, A., Simon, H.A.: Computer science as empirical inquiry: Symbols and search. In: ACM Turing award lectures, p. 1975 (2007)

[17] Newell, A., Simon, H.A., et al.: Human problem solving, vol. 104. Prentice-hall Englewood Cliffs, NJ (1972)

[18] Renkl, A.: Learning from worked-examples in mathematics: Students relate procedures to principles. ZDM **49**(4), 571–584 (2017)

[19] Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 (2016)

[20] Richardson, G.P., Andersen, D.F., Maxwell, T.A., Stewart, T.R.: Foundations of mental model research. In: Proceedings of the International System Dynamics Conference. pp. 181–192. EF Wolstenholme (1994)

[21] Staggers, N., Norcio, A.F.: Mental models: concepts for human-computer interaction research. International Journal of Man-machine studies **38**(4), 587–605 (1993)

[22] Suffian, M., Graziani, P., Alonso, J.M., Bogliolo, A.: FCE: Feedback based counterfactual explanations for explainable AI. IEEE Access **10**, 72363–72372 (2022). https://doi.org/10.1109/ACCESS.2022.3189432

[23] Suffian, M., Khan, M.Y., Bogliolo, A.: Towards human cognition level-based experiment design for counterfactual explanations. In: Mohammad Ali Jinnah University International Conference on Computing (MAJICC). pp. 1–5. IEEE (2022)

[24] Sweller, J.: The worked example effect and human cognition. Learning and instruction **16**, 165–169 (2006). https://doi.org/10.1016/j.learninstruc.2006.02.005

[25] Tsiakas, K., Barakova, E., Khan, J.V., Markopoulos, P.: BrainHood: towards an explainable recommendation system for self-regulated cognitive training in children. In: Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments. pp. 1–6 (2020)

[26] VanLehn, K.: Cognitive skill acquisition. Annual review of psychology **47**(1), 513–539 (1996)

[27] Varnum, M.E., Grossmann, I., Kitayama, S., Nisbett, R.E.: The origin of cultural differences in cognition: The social orientation hypothesis. Current directions in psychological science **19**(1), 9–13 (2010)

[28] Villaronga, E.F., Kieseberg, P., Li, T.: Humans forget, machines remember: Artificial intelligence and the right to be forgotten. Computer Law & Security Review **34**(2), 304–313 (2018)

[29] van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence **291**, 103404 (2021)

[30] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech. **31**, 841 (2017)

[31] Wu, J.Y., Yang, C.C., Liao, C.H., Nian, M.W.: Analytics 2.0 for precision education. Educational Technology & Society **24**(1), 267–279 (2021)

[32] Zhang, B.T.: Cognitive learning and the multimodal memory game: Toward human-level machine learning. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 3261–3267. IEEE (2008)