Dr hab. Przemysław Spurek, prof. UJ
Jagiellonian University
Faculty of Mathematics and Computer Science

REVIEW of PhD dissertation of **Spyridon Mouselinos**, MSc

Visual and Language Reasoning in Deep Learning Models

in the field of Natural Sciences in the discipline of Computer Science

The scientific achievement submitted by Spyridon Mouselinos, MSc is a collection of publications entitled *"Visual and Language Reasoning in Deep Learning Models."* It constitutes a sequence of thematically related scientific articles published in the peer-reviewed proceedings of international conferences. The thesis contains three scientific papers published in the post-conference proceedings of the prestigious ICLR, ACL, and EMNLP conferences.

The total number of MEiN points assigned to the publications included in the dissertation is 600, with each paper receiving 200 points. These works have also received several citations. According to Google Scholar, publication [P1] has 3 citations, [P2] has 11 citations, and [P3] has 15 citations. All three papers were published at top-tier international conferences (ICLR, ACL, and EMNLP), where Mr. Spyridon Mouselinos is listed as the first author.

From a bibliometric perspective, the evaluation of the scientific achievement presented by Spyridon Mouselinos is therefore **very good and even outstanding for this stage of his academic career**. The conferences in which the works were published are internationally recognized and widely regarded as the most prestigious venues in the field.

Spyridon Mouselinos is not the sole author of any of the publications. Each paper has three co-authors: the doctoral candidate and his supervisors. In all publications, the doctoral candidate is identified as the person with the leading contribution. **The individual contribution of Mr. Mouselinos to the works included in the dissertation is clearly defined, dominant, and beyond doubt.**

The doctoral dissertation is well written and clearly structured. The author introduces all concepts necessary for understanding the included works and presents the relevant issues in a coherent and accessible manner. **The individual publications together form a consistent and thematically connected body of scientific work**.

The doctoral dissertation is structured in a highly **logical and transparent** manner, a characteristic often found in achievements based on a sequence of thematically linked articles. The introduction section, including synthetic summaries of the publications, effectively grounds the reader in the context of the core problem: the lack of robustness and the tendency of AI models to rely on superficial heuristics. This **clear guide** to the subject matter facilitates an easy understanding of the motivation underlying the research and directs attention toward the main goals of the thesis.

The arrangement of the individual chapters, which closely correspond to the three publications, allows the reader to transition smoothly from the diagnosis of reasoning failures in the visual domain (VQA) through an analogous diagnosis in code generation to the synthesis and proposed solution in complex geometric reasoning (multi-agent systems). Such a logical sequence not only demonstrates the **profound coherence** of the body of work but also highlights the doctoral candidate's ability to conduct

systematic research that evolves from identifying weaknesses to proposing innovative mitigation strategies. Reading the dissertation is fluent and satisfying, effectively engaging the audience in a discussion about the foundations of contemporary artificial intelligence.

In the first paper, "Measuring CLEVRness: Black-box Testing of Visual Reasoning Models," the candidate tackles the fundamental challenge of verifying whether advanced Visual Question Answering (VQA) models possess compositional reasoning skills or merely leverage superficial statistical correlations, a phenomenon known as the Clever Hans Effect. The core thesis established here is that VQA models fail when the visual input is minimally perturbed yet semantically correct.

The main methodological innovation, presented in the sections, is the design of a black-box adversarial testing procedure framed as a two-player zero-sum game. The method is independent of the target VQA model's architecture and internal parameters, granting it broad utility. The game involves an Adversarial Player, an agent trained with Reinforcement Learning, whose sole objective is to discover the minimal semantic manipulation of the scene that causes the VQA model to change its correct answer. Crucially, this manipulation is strictly constrained to operate on the scene's symbolic graph, ensuring that the resulting images are rendered by the original CLEVR engine. This guarantees that the adversarial examples are in-distribution and semantically plausible, preserving realistic physics like shadows and lighting.

The experimental findings demonstrate evidence of model fragility. Top-performing VQA architectures, which achieve near-perfect accuracy on the original CLEVR dataset (e.g., MDetr at 99.7%), suffered dramatic performance collapses when evaluated against these semantically valid adversarial examples (with MDetr accuracy dropping to 60.0%). The central conclusion of the work is that failure is more strongly correlated with the logical complexity of the question than with the degree of visual alteration. This confirms that VQA models exploit fragile statistical shortcuts rather than relying on robust, invariant logical mechanisms.

Two questions arise regarding this section. First, since the Adversarial Agent in [P1] produced in-distribution manipulations, could these be used for Adversarial Training to harden VQA models, and why authors prioritize diagnosis over this solution? Second, which is more effective: augmenting the input with structured context, or adversarial training alone, for defense against the Clever Hans effect?

This article provides a robust scientific foundation for the entire dissertation, successfully diagnosing a core vulnerability in modern AI and introducing an original, model-agnostic methodology that shifts the field from passive accuracy testing to active behavioral verification. The publication's acceptance at ICLR 2022 confirms the work's high originality and technical rigor.

In the second paper, "A Simple, Yet Effective Approach to Finding Biases in Code Generation," the candidate expands the Clever Hans Effect hypothesis to Large Language Models (LLMs) used for code generation. The core argument is that these systems, despite achieving high benchmark scores, often rely on superficial lexical cues (such as function names or keywords) inherited from their vast training corpora rather than demonstrating authentic algorithmic understanding of the problem. The objective is to create a general, black-box tool capable of identifying these cognitive biases within programming tasks.

The key methodological contribution is the introduction of the *Blocks of Influence* paradigm. This technique structurally decomposes any coding challenge input into three distinct, manipulable components: the *Name Block*, the *Description Block*, and the *Example Block*. Manipulation of the *Name Block* is designed to expose memorization effects, where the model resorts to copying snippets from its training data. Manipulations of the *Description Block* by removing redundant keywords reveal lexical preference bias, testing the model's reliance on specific tokens rather than inferring context. Finally, removing the *Example Block* exposes underlying deficiencies in the reasoning. The methodology

employs context-aware filtering to ensure that manipulations preserve the task's global semantics, guaranteeing that any omitted information remains logically inferable from the remaining context.

Empirical results across large models such as Codex, CodeGen, and Incoder confirm cognitive biases on benchmarks such as HumanEval and MBPP. Minimal transformations consistently caused significant performance drops. For instance, combining anonymization with full keyword removal proved the most challenging, resulting in an average performance decline of approximately 40%, demonstrating that models struggle to deduce context solely from examples. These findings suggest a predictable hierarchy of reliance on input sources by the models. The work further explores bias mitigation through adversarial fine-tuning, showing that augmenting training sets with these transformed inputs allows models to partially recover performance and better generalize.

This publication is a powerful extension of the dissertation's central thesis, successfully adapting the adversarial testing methodology to the code generation task. The "Blocks of Influence" framework is an original, computationally inexpensive diagnostic tool that effectively reveals the internal cognitive biases of large code-generation models, fully justifying its publication at ACL 2023.

While reliance on the Name Block is defined as a memorization bias, does punishing the model for leveraging function naming conventions undermine its essential utility to a human programmer, who relies on names like sort_list to carry strong semantic meaning? Second, given the diagnostic focus, did the experiments show a correlation between task difficulty (e.g., easy, medium, hard categories from HumanEval) and susceptibility to adversarial attacks, such as keyword dropping? How could your Blocks of Influence framework, proven on clean benchmarks, be adapted to code generated in real-world, unstructured scenarios?

In the third paper, "Beyond Lines and Circles: Unveiling the Geometric Reasoning Gap in Large Language Models" (EMNLP 2024), the candidate investigates the limitations of advanced Large Language Models (LLMs) and Vision-Language Models (VLMs) in handling constructive geometry problems. Building directly on the insights from the two preceding diagnostic papers, this work treats geometry as a highly convincing test that demands the integration of stepwise planning, spatial comprehension, and tool-driven logic.

The paper identifies that while LLMs perform well on textual and algebraic tasks, they exhibit significant limitations in spatial reasoning and constructive geometry. To address this, the candidate introduces a novel multi-agent framework designed to improve geometric reasoning through collaborative problem-solving. The system decomposes the complex task into specialized, collaborative roles: the *Natural Language Solver* provides high-level rationale and planning, while the *Geometric Tool Solver* translates these plans into precise tool commands (like drawing lines or circles). This specialization is complemented by Validators who provide continuous feedback and iterative corrections to reduce errors and hallucinations.

Key innovations in this approach include the Adaptive Few-Shot Mechanism, which dynamically selects the most relevant past solutions to guide the current problem-solving process and minimize the need for irrelevant prompting. Furthermore, the system incorporates Naming Neutralization, which replaces alphabetical labels (like A, B, C) with neutral placeholders (e.g., "X") to successfully eliminate biases that often lead models to generate excessive or redundant construction steps. The results demonstrate that this structured, multi-agent collaboration significantly enhances the robustness and accuracy of general-purpose LLMs and VLMs on complex tasks, such as the Euclidea benchmark, confirming that structured decomposition can overcome the narrow failures identified in earlier work.

This third paper provides a critical synthesis and a solution focus on the entire dissertation cycle. It moves beyond merely diagnosing deficiencies (as in P1 and P2) to presenting a viable, model-agnostic strategy for achieving robust, nuanced reasoning by leveraging specialization and structured knowledge. The effectiveness of this multi-agent structure was also shown to generalize to other

mathematical domains (GSM8K, SVAMP). The work's originality and contribution to integrating advanced reasoning and multi-modal contexts are underscored by its publication at EMNLP 2024.

The demonstrated success of the multi-agent framework in geometric reasoning raises a crucial architectural question: Does this performance implicitly suggest a fundamental limitation in current monolithic (end-to-end) LLM architectures for solving complex, multi-modal tasks? In other words, do you conclude that achieving truly robust and structured reasoning requires the injection of explicit collaborative structures and external, specialized agents? To what extent does the performance of your multi-agent framework depend on the manual design and fine-tuning of agent roles and interaction protocols?

In conclusion, I would like to emphasize that the papers in this dissertation clearly demonstrate Spyridon Mouselinos's high-level scientific skills. His work reflects a deep and well-developed understanding of the subject matter and indicates strong potential for continued scientific development. It is also important to note that all three publications were accepted at leading international conferences widely regarded as top-tier in the field of artificial intelligence.

The doctoral dissertation meets the conditions set out in Art. 187 of the Act of July 20, 2018 Law on Higher Education and Science (Journal of Laws of 2024, items 1571). Moreover, in my assessment, the dissertation not only satisfies but clearly exceeds all customary and statutory requirements for doctoral dissertations. It constitutes an original and significant solution to well-defined scientific problems, demonstrates the candidate's broad and solid knowledge in technical informatics, and provides clear evidence of his ability to conduct independent, high-quality scientific research. Therefore, I recommend that Mr. Spyridon Mouselinos be admitted to the next stages of the doctoral procedure.

The dissertation further stands out for its originality, depth, and scientific maturity, representing a level of quality rarely observed at the doctoral stage. I strongly recommend that the dissertation be awarded a distinction.