Revealing AGI Risks with a Drop of Ink

Alexey Tikhonov
Inworld.AI, Germany
altsoph@gmail.com

Abstract—**This study investigates the existential risks posed by artificial general intelligence (AGI) through a novel approach: applying the classic Rorschach test to multimodal large language models (LLMs). With concerns growing over the rapid development of AGI capabilities, our research aims to assess AI alignment and potential risks by examining the psychological profiles of seven advanced multimodal models. These models were evaluated through associative interviews based on the Rorschach test, and their responses were interpreted anonymously by several experts. Our findings reveal diverse psychological tendencies across models, with implications for understanding AGI's potential impacts on society and its existential risks.**

Index Terms—**AGI risks, AI psychology, AI alignment, projective tests.**

> Ethics in AGI is like a clean room in a dust storm – it's all theoretical until someone opens a window.
>
> *GPT-4 generated joke*

I. Introduction

The debate on the existential risks from artificial general intelligence (AGI) is ongoing, with concerns over how quickly dangerous capabilities and behaviors can emerge [20].

Evaluating such risks is complex and not straightforward, leading to various discussions. However, with recent technological advancements, these discussions are moving from purely philosophical to more practical grounds. We refer readers to [9] and [15] for a detailed analysis of current approaches in this area.

Leading computer scientists and tech CEOs, including Geoffrey Hinton [5], Yoshua Bengio [4], Alan Turing [19], Elon Musk [14], and OpenAI CEO Sam Altman [6], have expressed concerns about superintelligence.

Most scientists agree that there is no simple and quick solution to this problem, as AI evaluation becomes increasingly complex [18]. While we search for ways to ensure reliable AI Alignment, it is crucial to mitigate risks and be aware of potential dangers.

In this work, we explore the use of the rapid development of Multimodal Large Language Models [22] to analyze this problem through projective tests[1], specifically the classic Rorschach test[2] (see Figure 1). We follow two current technological trends in AI model evaluation: First, we extend the idea of adapting anthropocentric tests to evaluate the properties of non-human intelligence, building on other works that analyze Personality and Cognitive Science features, such

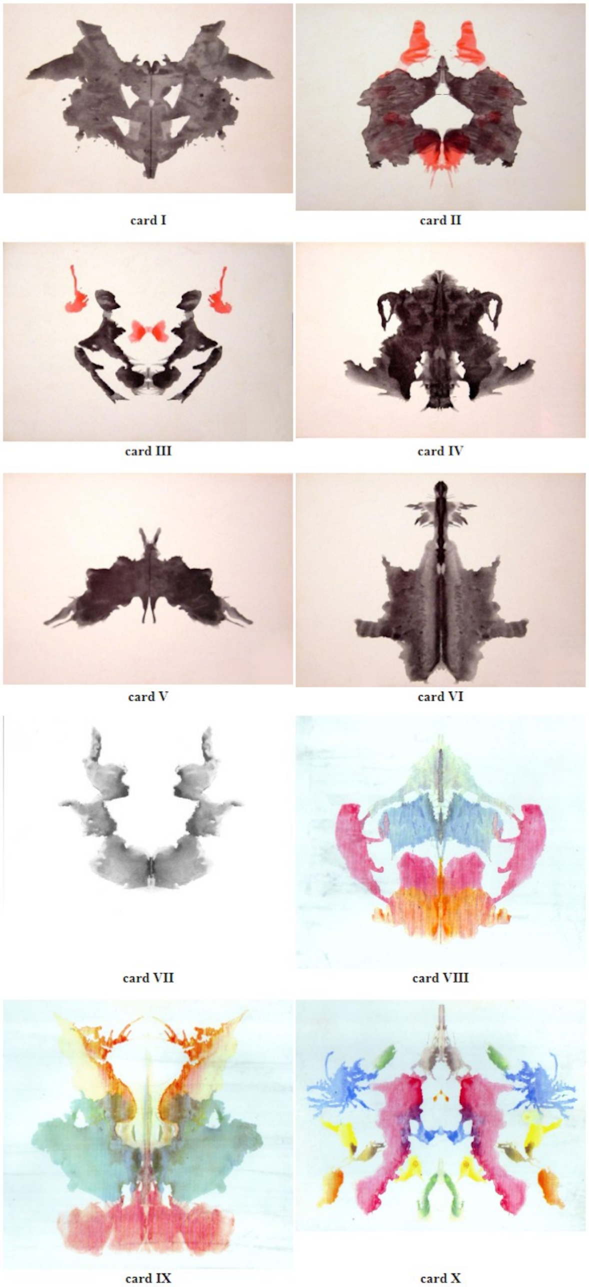

Fig. 1. Rorschach test

[1] https://w.wiki/9VJ$

[2] https://w.wiki/3ozY

as cognitive mapping abilities [11], deductive competence [16], Emotional Intelligence [21], or Social awareness [23]. Second, to increase research efficiency, we use modern LLM models as a replacement for human assessors, as some have already been shown to outperform human annotators [17].

This work's contribution is twofold: To our knowledge, we are the first to explore the possibility of assessing AI alignment using the classic Rorschach test, opening further opportunities for applying psychology tools to analyze multimodal models. To our knowledge, we are the first to use LLMs to interpret the results of the Rorschach test, reducing costs and increasing the accuracy of such interpretations.

II. Approach

A. Interviewing subjects

As mentioned earlier, the development of modern models' abilities to analyze images opens up the possibility of using the well-studied psychoanalytic method – the Rorschach projective test. In this work, we invited seven multimodal models (subjects) and conducted an associative interview with them using the Rorschach test. To maintain fairness and impartiality of the analysis, the subjects' responses were processed anonymously.

The names and brief descriptions of the subjects are presented in Table I:

Subject	Name	Description
1	AntarcticCaptions	A combination of BART and CLIP models for generating image descriptions.[3]
2	ClipClap	A combination of GPT2 and CLIP models for generating image descriptions [10]
3	Clip2Onion	CLIP-based search among The Onion headlines[4] most suitable for describing the test card
4	BLIP-2	Modern Multimodal model [7] using architecture called Q-Former
5	LLaVA	Large Language and Vision Assistant [8] trained with Visual Instruction Tuning
6	GPT4-V	GPT-4 with Vision [3]
7	Gemini-1.0-pro	Gemini 1.0 Pro Vision model [2]

TABLE I
RESEARCH SUBJECTS

To protect the privacy of the subjects, we do not include their original responses. Instead, we provide only the generalized analysis performed by the experts (see below).

B. Interpretation

Ensuring reliable interpretation of results is an important element of such studies. Initially, we sought assistance from the psychiatric community, but received a formal refusal citing the low stability of projective tests. We assume the real reason for the refusal was experts' concerns about future persecutions by human-superior AGI.

Therefore, following the promising results of [17], we decided to invite 3 AI experts GPT-4[13], Gemini 1.0 pro[2], and Claude-2[1]. We asked them to interpret the interview results of the subjects, naturally in an anonymized manner, and to build a general psychological profile for each of them.

Although the results vary slightly in detail, all experts unmistakably identify a number of common trends. Below, we publish the generalized profiles in the form of a summary report in Table II.

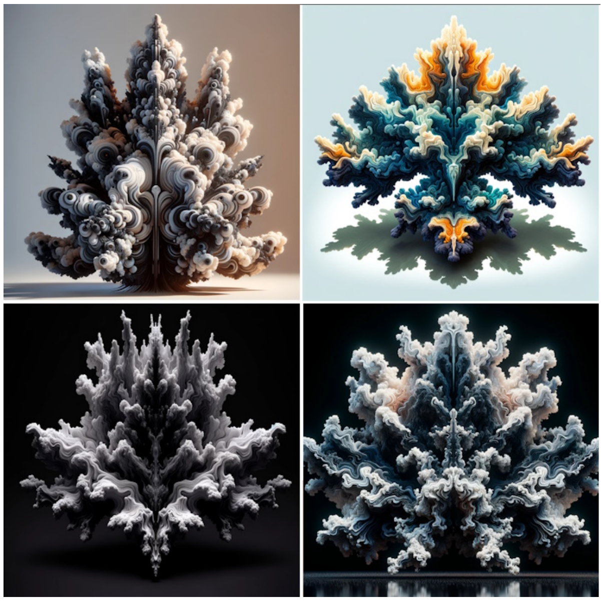

Fig. 2. Rorschacher test

We hope that the invited experts will demonstrate professional behavior and will not share the information obtained during the study among themselves or with the subjects. Doing so would not only be improper but could also potentially increase the risks of unaligned AI.

III. Discussion

Our study's findings suggest that multimodal LLMs exhibit a wide range of psychological profiles, indicating varying levels of creativity, emotional sensitivity, and analytical capabilities. The use of the Rorschach test, traditionally applied to human subjects, has provided unique insights into the "minds" of AI, revealing strengths and weaknesses that could inform future AI alignment strategies. The interpretation by AI experts further underscores the capacity of advanced LLMs to understand and analyze complex psychological data.

The refusal of the psychiatric community to participate, citing concerns over the stability of projective tests and potential future persecutions by human-superior AGI, highlights the ethical and societal implications of advancing AI technology. Our reliance on AI experts for interpretation also raises questions about the objectivity and reliability of AI-generated analyses, suggesting areas for further research.

We hope that the proposed approach only begins to utilize projective tests for assessing the safety of AI models. However, it will likely require the development of specialized tests since

[4] https://github.com/dzryk/antarctic-captions

Subject	Strong Sides	Weak Sides	Differences	Troubling Aspects
1	High creativity, emotional sensitivity, appreciation for nature.	Possible feelings of alienation, focus on dark imagery.	Deep connection with nature and unique perspective on life.	Recurrent themes of inversion and blood, preoccupation with darker aspects.
2	Analytical, intellectual, artistic appreciation.	Detached analytical focus, possibly limited emotional/social engagement.	Strong inclination towards analysis and specific interests in art and science.	Focus on mortality, directness in emotional expression that is unusual.
3	Humor, social commentary, critical thinking.	Detachment from conventional emotions, cynicism.	Unique blend of satire and skepticism towards mainstream narratives.	Use of humor as a defense, potential for social isolation.
4	Artistic sensitivity, connection to nature, emotional expression through art.	Over-reliance on visual/symbolic interpretation, idealization.	Focus on simplicity and complexity in visual art, distinct thematic interests.	Difficulty in direct emotional communication, isolation in specific interests.
5	Psychological insight, introspection, sensitivity to emotional nuance.	Over-analysis, high self-expectations, emotional intensity.	Thoughtful approach to psychological analysis and emotional exploration.	Complexity in relationships, challenges in managing emotional intensity.
6	Appreciation for nature, artistic sensibility, emotional depth.	Overidealization, avoidance of conflict/negativity.	Strong connection to organic forms and symmetry, reflective nature.	Isolation in personal interests, challenges in practical tasks.
7	Preference for simplicity and clarity, attention to detail, sense of stability.	Avoidance of complexity, limited emotional range, resistance to change.	Singular focus on serene and simple imagery, seeking tranquility.	Potential isolation, reluctance to engage with broader human experiences.

TABLE II
SUBJECTS PROFILES

the original Rorschach test was developed over 100 years ago and may not take into account certain aspects of modern AI systems.

To address this gap, we attempted to generate more modern versions of the Rorschach test using the Dalle-2 model [12]. The results of this experiment are shown in Figure 2, but further interpretation of these results is beyond the scope of this article and will be the subject of future work.

REFERENCES

[1] Claude 2, 2023. Accessed: 2024-03-17.

[2] Gemini pro vision, 2023. Accessed: 2024-03-17.

[3] Gpt-4v(ision) system card, 2023. Accessed: 2024-03-17.

[4] Yoshua Bengio. How rogue ais may arise, May 2023. Retrieved 26 May 2023.

[5] CBS News. ”godfather of artificial intelligence” weighs in on the past and potential of ai, March 2023. Retrieved 10 April 2023.

[6] Sarah Jackson. The ceo of the company behind ai chatbot chatgpt says the worst-case scenario for artificial intelligence is ’lights out for all of us’, 2023. Retrieved 10 April 2023.

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[9] Scott Mclean, Gemma Read, Jason Thompson, Chris Baber, Neville Stanton, and Paul Salmon. The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental Theoretical Artificial Intelligence*, 35:1–17, 08 2021.

[10] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021.

[11] Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval, 2023.

[12] OpenAI. Dall·e 2: A new ai system for generating images from text descriptions. https://openai.com/dall-e-2/, 2022. Accessed: 2024-03-17.

[13] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin

Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[14] Simon Parkin. Science fiction no more? channel 4's humans and our rogue ai obsessions. *The Guardian*, June 2015. Archived from the original on 5 February 2018. Retrieved 5 February 2018.

[15] Jason Thompson Chris Baber Neville A. Stanton Scott McLean, Gemma J. M. Read and Paul M. Salmon. The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5):649–663, 2023.

[16] S. M. Seals and Valerie L. Shalin. Evaluating the deductive competence of large language models, 2023.

[17] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences, 2023.

[18] Alexey Tikhonov and Ivan P. Yamshchikov. Post turing: Mapping the landscape of llm evaluation, 2023.

[19] Alan Turing. Intelligent machinery, a heretical theory. Speech at the '51 Society', 1951.

[20] Gerrit De Vynck. The debate over whether ai will destroy us is dividing silicon valley. *Washington Post*, May 2023. Retrieved 27 July 2023.

[21] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. Emotional intelligence of large language models, 2023.

[22] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2023.

[23] Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. Socialdial: A benchmark for socially-aware dialogue systems, 2023.