AwarePrompt: Using Diffusion Models to Create Methods for Measuring Value-Aware AI Architectures

Kinga Ciupinska[1][a], Serena Marchesi[1][b], Giulio Antonio Abbo[2][c], Tony Belpaeme[2][d] and Agnieszka Wykowska[1][e]

[1]*Social Cognition in Human-Robot Interaction (S4HRI), Italian Institute of Technology, Genova, Italy*

[2]*IDLab-AIRO – Ghent University – imec, Belgium*

kinga.ciupinska@iit.it, serena.marchesi@iit.it, giulioantonio.abbo@ugent.be, tony.belpaeme@ugent.be, agnieszka.wykowska@iit.it

Keywords: Awareness, Generative AI, Diffusion Models, Value-Aware AI, Insight, Neural Correlates

Abstract: The integration of diffusion models (DMs) into generative AI systems presents an approach with implications for ethical and moral AI development and our understanding of human-AI interaction. This study explores the intersection of generative AI, human values, and neuroscience, emphasizing the significance of value-awareness in AI systems. The methodology involves a behavioral experiment to evaluate the accuracy of DM-generated visual stimuli in capturing human values and related keywords. Results indicate promising match rates, marking stride in aligning AI systems with ethical and moral considerations. Additionally, the study introduces a criterion for selecting stimuli based on an "Aha" moment, setting the stage for an EEG experiment to explore the neural correlates associated with becoming aware of a value. This multidisciplinary study is a step toward the development of procedures to evaluate the effectiveness of Value-Aware AI systems in enhancing the perceived ethical and moral agency.

1 INTRODUCTION

Artificial Intelligence (AI) has become an integral part of our daily lives, with applications ranging from virtual assistants and recommendation systems to language generation. As AI systems, particularly generative language models, evolve and become more sophisticated, the importance of integrating value-awareness into their design and functionality becomes paramount. Value-awareness refers to the incorporation of ethical considerations, cultural sensitivity, and a deep understanding of human values into AI systems. In the context of generative AI, which involves the creation of text, the implications of value-awareness are far-reaching and crucial for responsible and ethical AI development.

Generative AI, particularly in language models like OpenAI's GPT-3, has demonstrated remarkable capabilities in understanding and generating text. Diffusion Models (DMs) for computer vision are trained to iteratively remove noise from an image that was blurred. The trained models can then be used for multiple applications including image generation, image manipulation (adding or removing elements from the image, changing the lighting, etc.) and image improvement – denoising, increasing the resolution (Croitoru et al., 2023; Liuand et al., 2022). Access to these models is increasing constantly, and more people can use them for a wide range of tasks: from improving a rushed selfie to generating graphics for slides and publications, from new kinds of digital art, to very convincing images to show side by side a fake news article generated using a Large Language Model.

Given the strong impact on humans' lives, the role of generative AI in decision-making processes has prompted increased scrutiny regarding the alignment of these systems with human values. While generative AI has the potential to acquire additional knowledge from the real world, incorporating human value judgments into these systems remains a significant challenge, and there are few specific strategies identified for addressing this issue (Hendrycks et al., 2021). To effectively collaborate with people and navigate hu-

[a] https://orcid.org/0000-0002-9909-4400
[b] https://orcid.org/0000-0001-9931-156X
[c] https://orcid.org/0000-0001-6301-0028
[d] https://orcid.org/0000-0001-5207-7745
[e] https://orcid.org/0000-0003-3323-7357

man environments, AI systems must have the ability to understand, interpret, and predict human decisions. Many human decisions involve moral considerations such as concern for harm, justice, fairness (Turiel, 1983), or broader issues of interdependent rational choice (Hayakawa, 2000), reflecting the intricate interplay between individual ethical frameworks and societal values that shape the complex landscape of human behavior.

The adaptive capacity of the human moral mind allows individuals to cooperate for mutual benefit in changing circumstances and emerging opportunities to both help and harm (Tomasello and Vaish, 2013). However, predicting human moral judgment poses a formidable challenge for AI systems, especially when it comes to responding sensibly to novel situations that do not arise during training (Hendrycks et al., 2021). This flexibility, central to human moral cognition, poses a particularly complex challenge for AI systems.

As we have already mentioned, the ethical implications of algorithmic decision-making and the synchronization of artificial intelligence systems with human values have become major focal points in the field of AI. Therefore, in recent years, several attempts have been made to create moral AI systems (Charisi et al., 2017; Gonzalez Fabre et al., 2021). These methodologies aim to effectively communicate moral and ethical values in a way that both machines and humans can understand. Therefore, drawing inspiration from the VALUENET dataset, a comprehensive repository of human-driven dialogues centering on values (Qiu et al., 2022), we embark on a journey to visually articulate complex ethical and moral concepts. The chosen value-related words, meticulously curated from this dataset, form the foundation for generating a diverse array of images through DMs. By employing DMs, we can systematically generate visual stimuli that encapsulate a spectrum of ethical and moral values. This process is vital for evaluating the fidelity of generative AI systems in reflecting the nuances of human values.

The first goal of using DMs in our research was to assess how well AI systems represent human values and related words. Through a careful selection of input data and fine-tuning of model parameters, we were able to generate diverse visual stimuli that encapsulate a wide range of ethical considerations. These stimuli were then presented to human subjects for evaluation, allowing us to gauge the alignment between the generated content and human values. This iterative process provides valuable insights into the strengths and limitations of generative AI systems in capturing the complexity of ethical and moral frameworks.

Building on the success of DMs in generating ethically relevant visual stimuli, the second goal involves leveraging these stimuli to study the neural correlates of becoming aware of a value. The "aha moment" or insight research, which focuses on sudden realizations and shifts in awareness (Sprugnoli et al., 2017), can be applied to the domain of ethical decision-making. By carefully selecting stimuli that evoke ethical insights, we can further conduct EEG experiments to identify the neural processes associated with the recognition and internalization of values.

Therefore, the implications of our research extend to the realm of human-AI interactions and the development of value-aware AI architectures. By analyzing the effectiveness of visual stimuli generated by DMs, we aim to contribute to the understanding of an AI ethical framework that responds to human perspectives and sensitivities. This work focuses on developing three important points:

1. Advancements in Value Representation This research is expected to contribute to the understanding and development of methods for representing moral and ethical values in AI systems. The combination of linguistic and visual stimuli generated by DMs provides a comprehensive approach to encapsulate the diverse range of human values and related keywords.

2. Validation of Representations The image recognition experiment serves as a critical validation step, ensuring that the generated visual stimuli appropriately convey the intended values. This step is vital for the practical application of AI systems in contexts where ethical and moral decision-making is essential.

3. Neural Correlates of Becoming Aware of Value Such AI-generated stimuli will serve as tools that allow for systematically exploring the neural underpinnings of understanding ethical and moral values. The conclusions obtained from this research will not only contribute to understanding how ethics and moral-related values are processed at the neural level but also will be used to develop behavioral and neuroscience (EEG) studies to assess the success of value-aware AI systems in increasing perceived moral agency.

2 METHODOLOGY

2.1 Participants

In order to test how well people will recognize values represented by stimuli generated using DMs, a total

of 36 participants (12 for each of 3 versions, see Section 2.2.3) completed a pilot behavioral experiment via Prolific (20 males, age 20-54, M = 33.33, SD = 9.69). All participants gave informed consent prior to the experimental session and were compensated for their time (9 EUR per session).

2.2 Materials and Experimental Paradigm

2.2.1 Selection of Values

The selection of values was based on the VALUENET dataset (Qiu et al., 2022), a dataset designed for human value-driven dialogue systems. The dataset encompasses a wide range of moral and ethical values expressed through human conversations. From this dataset, a list of words representing key values was curated, forming the basis for generating stimuli through DMs. Our glossary contains all the words that were listed in Figure 1 of Qiu and collaborators (2022). To keep our experiment as consistent as possible, all words were transformed to their noun form (i.e. if a word was an adjective in the original text, we changed it to a noun). The study was conducted on people whose native language is Italian. The word list was translated by people whose native language is also Italian and English is their second language.

2.2.2 Generation of Visual Stimuli Using DMs

The images are generated using Stable Diffusion XL (Podell et al., 2023), a text-to-image model. In particular, the setup comprises a base model, `stable-diffusion-xl-base-1.0`, which takes a text prompt and produces an intermediary output. This is then elaborated by a refiner model, `stable-diffusion-xl-refiner-1.0`, to obtain the final image. In addition, the model allows to specify negative prompts, containing a description of what should be avoided in the picture.

After several tests, the final prompt consists of two parts: a noun (the name of the value) and a definition taken from the Oxford Dictionary. The choice of including the definition was dictated by two reasons: to help disambiguate those terms that have multiple meanings, and to add context to the model prompt. Indeed, text-to-image models work best when provided with a clear description of the desired image. In this case, adding the definition of the terms gives higher quality results compared to using only one word. An example of a prompt is the following: *"Wisdom: capacity of judging rightly in matters relating to life and conduct; soundness of judgement in the choice of means and ends."*

However, a test run over a subset of the values revealed three main issues with the images. First, this method produced images containing text, and (secondly) the style of the images resembled the etchings commonly found in books. One explanation for both of these issues is the formal register of the dictionary descriptions, which often are associated with these image features. The third problem is that, in some cases, the images produced were chaotic and unclear, whereas for the scope of this evaluation it is best if the images have a single subject. For these reasons, a negative prompt was added, which works similarly to a normal prompt, except that it specifies what is *not* desired in the output image. The negative prompt contained two terms to reduce each of the three problems, for a total of six words separated by a comma. We obtained good results with the following negative prompt: *"text, letters, drawing, scan, complex, chaos"*.

The implementation of a negative prompt facilitated the generation of stimuli wherein each image encapsulates a distinct scene as opposed to a composite amalgamation of diverse elements. The employment of this negative prompt was principally guided by a strategic consideration for subsequent utilization of these stimuli in EEG testing. The rationale underlying the adoption of stimuli featuring solitary scenes stems from the endeavor to minimize potential confounds during EEG measurements. The incorporation of intricate and disorderly visual stimuli can introduce extraneous noise in the recorded signal, thus necessitating the adoption of simplified and focused visual stimuli to enhance signal fidelity and interoperability.

For each of the values, three images were generated using 50 inference steps with an 80% ratio (40 for the base model and 10 for the refiner), the guidance scale used was 7, which is a commonly used value to obtain quality results that do not stray away from the prompt. Figure 1 shows an example of the images generated for the term *wisdom*, using the prompt reported above[1].

Figure 1: The three images generated for the value *Wisdom*.

[1] All stimuli generated for this study are available at https://doi.org/10.5281/zenodo.10516944.

2.2.3 Experimental Design

To validate the effectiveness of the generated stimuli in representing values, an image recognition experiment was conducted. Our experimental procedure was based on paradigms examining the *Aha! effect* (Sprugnoli et al., 2017). We decided to base our paradigm on procedures examining the insight moment because we considered it an appropriate measure that would allow us to find neural markers of becoming aware of a value in further EEG experiments.

To check whether DMs can adequately generate images representing values and to select a set of appropriate stimuli for further EEG experiments (described in Section 4), a pilot behavioral experiment was conducted. The experimental procedure was similar to that adopted by previous studies (Zhao et al., 2014; Sandkühler and Bhattacharya, 2008; Sheth et al., 2009) (see in Figure 2).

The experimental procedure was written in PsychoPy version 03.02.2021 (Peirce, 2007). Since our stimuli set contained visual representations of 162 values, we decided to split it into 3 smaller sets (54 stimuli in each). Stimuli for each set were selected randomly. Each participant was assigned to one of the three versions also randomly. At the beginning of the experiment, participants were provided with training consisting of 5 stimuli. On each trial, a white fixation cross (10 x 10 pixels) was displayed centrally on a gray background for a random duration between 2 and 3 s. Next, a set of 3 images was presented on the screen for 60 s (coordinates were: -300, 0; 0, 0; 300, 0 pixels). Participants were required to come up with an appropriate word that was represented by the stimuli. During this initial period, the participants were instructed to press the space bar if they came up with the solution before the time was up. Next, participants were asked to enter the word or leave the field blank if they could not come up with any idea. After that, they were asked to provide the following subjective ratings: (1) the Rating of Suddenness: the degree of suddenness of the emergence of the answer ranging from 1 (the task was solved step by step) to 4 (the answer came very sudden). (2) the Rating of Confidence: the degree of confidence the participants felt about the answer they reported before, which also ranged from 1 (no confidence) to 4 (full of confidence). (3) the Rating of Restructuring: the restructuring process was described as rejecting the original meaning of a word and reinterpreting it in a new way, ranging from 1 (no restructuring) to 4 (full restructuring). After these subjective ratings, the correct answer was presented on the screen for 3 s. Next, participants were instructed to report their insight or "aha" feeling described as the moment when they suddenly understood or discovered something that was previously unclear or difficult to understand. In the case of this task, it means that they suddenly understood the correct answer because they didn't know/understand it before, ranging from 0 (no insight feeling) to 1 (having insight feeling). Finally, they were asked to indicate on 5-point Likert scale (Likert, 1932) how strongly they agree that a given set of pictures represents a given word properly (from 1 - strongly disagree to 5 - strongly agree). Before each rating and the presentation of a correct answer, a blank was presented for a random duration between 500 and 1000 ms. The whole experimental session lasted approximately 1 hour.

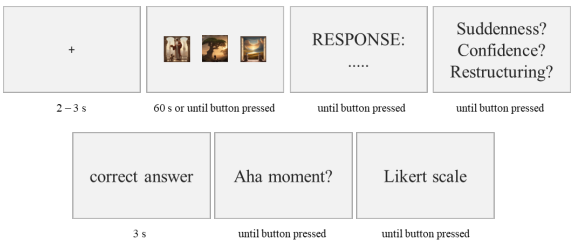

Figure 2: The flow map of the experiment.

Accuracy and the Likert scale (as an objective and subjective measure, respectively) were used to check whether DMs correctly represent human values (i.e., how accurately people recognize the value represented and how they evaluate the appropriateness of the images' representation of this value). The "Aha Rating" was used as a measure that allows the selection of appropriate stimuli for the EEG experiment, because in experiments examining insight, the "Aha" needs to be a sudden awareness or an unexpected comprehension of an answer to a question, not obvious before (Shen et al., 2013). The suddenness, confidence, and restructuring ratings were used as additional measures to identify the classical features of these insight problems (Zhao et al., 2014).

2.3 Data Analysis

Data analysis was conducted using R Studio version 2022.07.1. To check how well a given set of stimuli represents value, we analyzed accuracy and rating on the Likert scale. Accuracy was calculated based on participants' responses i.e., when a word indicated by participants matched the word of a value the answer was correct (1), otherwise (participant's answer did not match the word or no answer was provided) the answer was incorrect (0). For subjective rating, we dichotomized the 1-5 Likert scale scores as follows: 1, 2, and 3 ratings indicated low agreement that a set of stimuli represents a given value correctly (0), while

4 and 5 indicated high agreement (1).

The feeling of "Aha moment" indicated directly by participants after seeing the correct answer was a direct measure of the insight. Ratings equal to 0 indicated no insight feeling, while answers equal to 1 showed having insight feeling. Additionally, previous research (Zhao et al., 2014; Sandkühler and Bhattacharya, 2008; Sheth et al., 2009) defined insight as a solution accompanied by feelings of high suddenness, high confidence, and high restructuring. Based on this operationalization, to assess the proportions of insightful and non-insightful solutions we dichotomized the 1-4 scale scores on each component, as follows: 1 and 2 ratings indicated low suddenness, confidence and restructuring (0), while 3 and 4 indicated high scores (1). Thus, stimuli for which participants indicated feelings of insight (i.e., average rating higher than 0.5), and additionally high suddenness (i.e., average rating higher than 2), high confidence (i.e., average rating higher than 2), and high restructuring (i.e., average rating higher than 2) will be selected as a set for the further EEG experiment. Noninsightful stimuli were indicated by any other combinations of the components' levels, e.g., no insight feeling, low suddenness, low confidence, and/or low restructuring.

3 RESULTS

3.1 Validation of the Representation of the Value by Generated Stimuli

Data showed that the average objective accuracy rate reached to 15% (SD = 0.16), and only 17 out of 162 (10%) stimuli sets were correctly recognized in more than 50% of trials (M = 0.63, SD = 0.12). For the subjective rating measured by the Likert scale, there were 85 (52%) stimuli sets in which the rating values were on average larger than 3 (M = 3.67, SD = 0.47; meaning that participants agreed that a given set of images accurately represents a value). See Table 1 for the summary of the results.

3.2 Validation of the Insight Feeling

Our data indicated that 76 stimuli sets (47%) were accompanied by the feeling of insight (M = 0.69, SD = 0.11), and additionally high suddenness (M = 2.69, SD = 0.27), high confidence (M = 2.59, SD = 0.24), high restructuring (M = 2.69, SD = 0.26). The remaining 86 stimuli sets (53%) were not accompanied by the feeling of insight (M = 0.34, SD = 0.07). However, they were also accompanied by high suddenness (M = 2.38, SD = 0.39), confidence (M = 2.49, SD = 0.49), and restructuring (M = 2.03, SD = 0.22). See Table 2 for the summary of the results and Figure 3 for results visualization.

The list of 76 values that will be used for the EEG experiment looks like this: accomplishment, antiquity, authority, beauty, brilliantness, challenge, charity, Christian, classic, cleanliness, comfort, communication, compassion, compatibility, completion, conscience, courage, creation, devoutness, discipline, divinity, equality, eternity, excitement, exercise, exploration, faithfulness, force, forgiveness, formality, friendship, fun, generosity, gentleness, guard, health, helpfulness, humanity, indulgence, intelligence, intensity, interests, Islam, kindness, leadership, limitlessness, loyalty, manner, mercy, norms, order, orthodoxy, parents, participation, passion, peace, piousness, principle, production, protection, regulation, relax, republicanism, rich, rights, safekeeping, satisfaction, self-reliance, sociality, sovereignty, spirituality, strictness, support, unity, wisdom, work.

Table 1: Statistics for assessing how well generated visual stimuli sets represent values.

	M	SD
Accuracy (0/1)	0.25	0.12
Likert's rating (1-5)	3.67	0.47

Table 2: Statistics of *Aha!* Ratings.

	Insight		No Insight	
	M	SD	M	SD
Insight feeling (0/1)	0.69	0.11	0.34	0.07
Suddenness (1-4)	2.69	0.27	2.38	0.39
Confidence (1-4)	2.59	0.24	2.49	0.49
Restructuring (1-4)	2.69	0.26	2.03	0.22

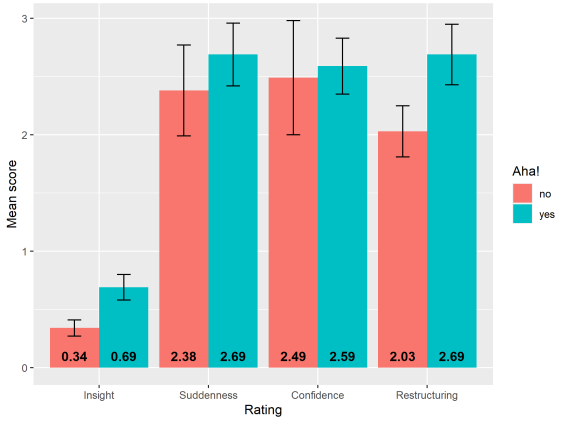

Figure 3: Comparison of means on ratings measuring insight feeling for stimuli accompanied by the feeling of Aha! and not.

4 DISCUSSION

In this study, we based our experimental paradigm on traditional 'Aha! effect' research and used diffusion models to create visual stimuli representing words related to human values. The work presented focused on two primary goals: evaluating how well generative AI systems represent such moral and ethical-related words and selecting stimuli suitable for studying the neural correlates of becoming aware of a value.

4.1 Alignment with Human Values

The results of the behavioral experiment shed light on the effectiveness of DM-generated visual stimuli in representing human values. Results showed that our prompt did not generate stimuli that represent values that can be objectively recognized by participants. However, the subjective measure of compliance (i.e., Likert scale) provides a comprehensive view, indicating a noteworthy degree of accuracy in conveying ethical and moral values by visual stimuli. While it is clear that there is much work to be done, positive compliance rates (measured by Likert scale) suggest that DMs can effectively capture and communicate a diverse range of values. The low accuracy of recognizing a given value based on generated stimuli may be the result of linguistic limitations, such as the existence of synonyms. The human language is very complicated and we can often find several words (Turney, 2008) for one expression. Therefore, it may be that our subjects, even if they understood what concept was represented by the series of DMs-generated stimuli, used a synonym for the correct word. Therefore, one of our future plans will also be to check whether subjects actually used synonyms to describe a given set of stimuli. It should also be noted that the values found in the glossary refer to abstract concepts that are difficult to grasp and visualize, and therefore it may be difficult to label them correctly.

It is crucial to highlight that a fundamental drawback exists in these models due to the initial bias present in their training sets. The representation of different values may still be somewhat homogeneous in Western society. As noted by (Henrich et al., 2010) a significant portion of research in human psychology operates under the assumption that fundamental cognitive processes are universally shared, and findings from one population can be universally applied. As the authors note, most of the research results and assumptions are based on the Western population.

In sum, behavioral findings suggest that the inability to generate stimuli that participants could objectively recognize as values prompts a critical reflection on the methodology and the underlying assumptions. It suggests that the chosen method or prompt might not effectively capture or convey the nuanced and subjective nature of human values. This recognition is crucial for refining future experimental designs and prompts, taking into account the complex and multifaceted nature of values. While the negative result may alter the anticipated trajectory of the study, it catalyzes further exploration, refinement, and adaptation of methodologies.

Beyond their role in eliciting the "Aha moment" during EEG experiments, these images hold potential in realistic scenarios. They can serve as educational tools, enriching learning materials and presentations to foster a more profound understanding of abstract values. Moreover, the validated stimuli, especially those associated with the insight feeling, may find utility in psychological and therapeutic settings, aiding introspection and discussions on personal values. In the realm of design and creativity, these images may inspire artistic endeavors, offering a visual language to explore and communicate complex societal themes. However, the variability in participant responses underscores the need for a nuanced approach in future applications, acknowledging diverse interpretations to maximize impact across different contexts.

4.2 Potential for Value-Aware AI Development

As we described earlier, in times of extraordinary development of AI systems, their proper interaction with human agents is extremely important. Our study's findings offer practical implications for the ethical development of AI systems, as the ability of generative AI to create visual stimuli that align with human values is a crucial step towards responsible AI. Firstly, by ensuring that AI-generated content reflects human values, we mitigate the risk of unintentionally perpetuating biases or promoting content that may be ethically questionable (Chimbga, 2023). Secondly, value-aligned visual stimuli can enhance user trust and acceptance of AI applications. Users are more likely to engage positively with technology that resonates with their values, fostering a sense of reliability and ethical responsibility (Ma and Huo, 2023). Thirdly, in scenarios where AI systems interact with users, generating content in line with human values contributes to more ethical and respectful interactions (Dignum, 2017).

Regardless of the application, AI must consider societal values, ethical concerns, and moral considerations (Dignum, 2017). This is mainly because AI

systems are tools under the control of human users, their potential autonomy and learning capabilities necessitate a deliberate incorporation of accountability, responsibility, and transparency principles in the design process (Charisi et al., 2017). We propose that AI development should prioritize the consideration of human values in AI systems because emphasizing value-aware AI performance can lead to the exploration of innovative techniques and applications.

The success and challenges observed in conveying values-related words through visual stimuli emphasize the need for tailored approaches in designing awareness architectures. Future architectures should prioritize adaptability to individual cognitive processes, acknowledging the nuanced ways individuals interpret and resonate with visual representations. Incorporating mechanisms that account for the variability in cognitive responses can enhance the effectiveness of awareness-building initiatives. Integrating such insights into the design of awareness architectures can potentiate their impact, fostering a more profound and meaningful understanding of values in diverse contexts.

4.3 Insightful Selection for EEG Study

The identification of values that evoke an "Aha!" moment or insight feeling, accompanied by high suddenness, confidence, and restructuring, serves as a pivotal aspect of this study. This insightful selection criterion serves as a deliberate and strategic foundation for the subsequent EEG experiment, ensuring that the neuroscientific investigation delves into values-related words that evoke profound cognitive processes. Specifically, the sudden recognition of moral or ethical-related value, as indicated by the participants' heightened experiences of insight, becomes a focal point for understanding the neural underpinnings of value perception. Thanks to this we will develop experiments that examine how people perceive AI systems as moral and intentional agents. Such an approach will enable us to make an important step toward the neuroscientific assessment of the success of Value-Aware AI systems in increasing the perceived moral and ethical agency.

5 CONCLUSION

In conclusion, this study represents a significant step forward in the convergence of generative AI, ethical considerations, and neuroscience. The integration of diffusion models as a means to generate visual stimuli for representing ethical and moral values not only demonstrates the feasibility of aligning AI systems with human values but also opens up avenues for exploring the neural underpinnings of value awareness. As the field advances, the interplay between AI and human values, as studied through the lens of DMs, contributes not only to a deeper understanding of the cognitive processes involved in ethical decision-making but also will help us to test whether and to what extent people perceive AI systems as capable of understanding and making moral and ethical values. This multidisciplinary approach positions our study at the forefront of ethical AI development, offering valuable insights for researchers, developers, and policymakers navigating the complex intersection of technology and human values.

ACKNOWLEDGEMENTS

This work has received support from the European Union under the European Innovation Council (EIC) research and innovation programme, Project VALAWAI, Grant Agreement number 101070930.

REFERENCES

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., ..., and Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint arXiv:1703.04741.

Chimbga, B. (2023). Exploring the Ethical and Societal Concerns of Generative AI in Internet of Things (IoT) Environments. In Southern African Conference for Artificial Intelligence Research (pp. 44-56). Cham: Springer Nature Switzerland.

Croitoru, F. A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Dignum, V. (2017). Responsible artificial intelligence: designing AI for human values. ITU Journal: ICT Discoveries, Special Issue No. 1.

Gonzalez Fabre, R., Camacho Ibáñez, J., and Tejedor Escobar, P. (2021). Moral control and ownership in AI systems. AI & SOCIETY, 36, 289-303.

Hayakawa, H. (2000). Bounded rationality, social and cultural norms, and interdependence via reference groups. Journal of Economic Behavior & Organization, 43(1), 1-34.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. (2021). Unsolved problems in ml safety. arXiv preprint arXiv:2109.13916.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29-29.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of psychology.

Liuand, N., Liand, S., Duand, Y., Torralbaand, A., and Tenenbaumand, J. B. (2022). Compositional visual generation with composable diffusion models. European Conference on Computer Vision (pp. 423-439).

Ma, X. and Huo, Y. (2023). Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. Technology in Society, 75, 102362.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. Journal of neuroscience methods, 162(1-2), 8-13.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis.

Qiu, L., Zhao, Y., Li, J., Lu, P., Peng, B., Gao, J., and Zhu, S. C. (2022). Valuenet: A new dataset for human value driven dialogue system. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 11183-11191).

Sandkühler, S. and Bhattacharya, J. (2008). Deconstructing insight: EEG correlates of insightful problem solving. PLoS one, 3(1), e1459.

Shen, W., Liu, C., Yuan, Y., Zhang, X., and Luo, J. (2013). Temporal dynamics of mental impasses underlying insight-like problem solving. Science China Life Sciences, 56, 284-290.

Sheth, B. R., Sandkühler, S., and Bhattacharya, J. (2009). Posterior beta and anterior gamma oscillations predict cognitive insight. Journal of cognitive neuroscience, 21(7), 1269-1279.

Sprugnoli, G., Rossi, S., Emmendorfer, A., Rossi, A., Liew, S. L., Tatti, E., ..., and Santarnecchi, E. (2017). Neural correlates of Eureka moment. Intelligence, 62, 99-118.

Tomasello, M. and Vaish, A. (2013). Origins of human cooperation and morality. Annual review of psychology, 64, 231-255.

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge University Press.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. arXiv preprint arXiv:0809.0124.

Zhao, Q., Li, Y., Shang, X., Zhou, Z., and Han, L. (2014). Uniformity and nonuniformity of neural activities correlated to different insight problem solving. Neuroscience, 270, 203-211.