Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods

Thilo Hagendorff
thilo.hagendorff@iris.uni-stuttgart.de
University of Stuttgart
Interchange Forum for Reflecting on Intelligent Systems

Abstract – Large language models (LLMs) are currently at the forefront of intertwining AI systems with human communication and everyday life. Due to rapid technological advances and their extreme versatility, LLMs nowadays have millions of users and are at the cusp of being the main go-to technology for information retrieval, content generation, problem-solving, etc. Therefore, it is of great importance to thoroughly assess and scrutinize their capabilities. Due to increasingly complex and novel behavioral patterns in current LLMs, this can be done by treating them as participants in psychology experiments that were originally designed to test humans. For this purpose, the paper introduces a new field of research called "machine psychology". The paper outlines how different subfields of psychology can inform behavioral tests for LLMs. It defines methodological standards for machine psychology research, especially by focusing on policies for prompt designs. Additionally, it describes how behavioral patterns discovered in LLMs are to be interpreted. In sum, machine psychology aims to discover emergent abilities in LLMs that cannot be detected by most traditional natural language processing benchmarks.

Keywords – machine psychology, machine behavior, artificial intelligence, large language models, explainability

1 Introduction

Recent surges in computing power, data availability, and research on learning algorithms – especially deep neural nets – have yielded powerful AI systems that are used in almost every area of society. Among the range of different AI technologies, large language models (LLMs) are especially gaining more and more attention. By providing access to LLMs via easy-to-use graphical user interfaces, companies like OpenAI ushered in the success of models like ChatGPT or GPT-4 (OpenAI 2022, 2023) with millions of daily users. Moreover, LLMs are at the cusp of being deployed as search engines and virtual assistants in high-stakes areas, thus impacting societies at large. In other words, next to humans, LLMs are increasingly becoming important contributors to the infosphere, causing immense societal changes by normalizing communicative relations between humans and artificial systems.

This development calls for methods with which traits and behavioral patterns of LLMs can be investigated and evaluated. These methods comprise approaches to foster explainability, for instance by using attention score visualizations (Vig 2019); measures for bias detection and mitigation, for instance by looking at underspecified questions that discover stereotypes (Li et al. 2020); factual correctness improvements and hallucination avoidance, for instance by linking model outputs and source web pages (Menick et al. 2022); ethical impact assessments, for instance by predicting malicious use cases (Weidinger et al. 2022); and many more. However, due to the ability of LLMs to reason and to engage in open-domain conversations, new and further-reaching opportunities for investigating traits and behavioral patterns of LLMs become necessary.

This is where machine psychology comes into play. It aims to elicit mechanisms of decision-making and reasoning in LLMs by treating them as participants in psychology experiments. Decades of researching humans and "accessing" the human psyche via language-based tools resulted in a plethora of methods and test frameworks that can now be harnessed to investigate not just humans, but also artificial agents. Since the increase in capabilities in artificial neural nets, especially transformers (Vaswani et al. 2017), has led to an increase in technical opacity, too, it is promising to borrow these methods that were designed to investigate another opaque structure, namely biological neural nets. Instead of deriving properties of LLMs by scrutinizing their intrinsic properties, meaning their neural architecture, machine psychology adapts a behaviorist perspective. It focuses on the correlation between prompts (inputs) and prompt completions (outputs) when applying psychological tests. This way, machine psychology can acquire systematic insights into patterns of machine behavior (Rahwan et al. 2019) and emergent reasoning abilities (Wei et al. 2022a) in LLMs that in many cases cannot be discovered with traditional benchmarks.

This paper conceptualizes the nascent field of machine psychology. It first describes the state of the art by bringing together various studies that already utilize test frameworks developed in psychology to probe machine behavior in LLMs. However, as of yet, these studies lack both an umbrella term as well as common methodological guardrails under which they can operate. Hence, in the second section, the paper describes the many links that one can forge between different fields of psychology and machine behavior research, as well as the many open research questions that future studies can tackle. Furthermore, in the third section, the paper defines methodological rules that are central to the field of machine psychology, especially by focusing on policies for prompt design. The penultimate section discusses several pitfalls when interpreting an LLM's behavior by using rich psychological concepts and terms. The last section concludes the paper.

2 State of the art

Numerous benchmarks to investigate LLMs have been developed over the last years, with BIG-bench being the largest of them, comprising more than 200 tests (Srivastava et al. 2022). These benchmarks apply problems from linguistics, math, commonsense reasoning, physics, logic puzzles, etc. to LLMs. They also measure things such as humor detection, metaphor understanding, social biases, and many more.

Some of these benchmarks overlap with study objects that can also be designated to machine psychology, most notably investigations on social stereotypes in LLMs (Field et al. 2021; Sheng et al. 2019; Nadeem et al. 2020). However, while traditional benchmarking primarily aims at optimizing LLMs against the backdrop of specific metrics, machine psychology is not primarily interested in increasing an LLM's performance. In addition to that, compared to traditional natural language processing benchmarks that measure abilities such as reading comprehension, translation, numerical reasoning, paraphrasing, truthfulness, etc., machine psychology can tackle broader, more complex scenarios and identify high-level traits in individual LLMs as well as in multiple instances of an LLM interacting with each other. This is, among others, especially important for AI safety and alignment research (Amodei et al. 2017; Hendrycks et al. 2022; Hendrycks and Mazeika 2022), which tries to identify emergent properties such as goal setting, self-awareness, deceptiveness, or other potentially malevolent and unsafe behavior in AI systems that are indiscernible in a low-level analysis as well as largely overlooked in entrenched auditing frameworks (Mökander et al. 2023). Moreover, methods of machine psychology will become crucial when evaluating the transition from narrow AI to artificial general intelligence (AGI). Recent studies in this regard already stress their relatedness to traditional psychology research (Bubeck et al. 2023).

A concept akin to machine psychology is the investigation of "machine behavior". The seminal paper in this field primarily draws analogies between ethology, the study of animal behavior, and machine behavior research (Rahwan et al. 2019). This is due to the study's scope, which pertains to AI systems and robots in general and which in most cases cannot be studied by using natural language. However, since this does not hold for LLMs, language-based methods from psychology can augment or replace comparatively simpler methods from ethology. In this context, a first set of studies investigating LLMs by using means from psychology have been published, but without putting their efforts under an umbrella term like "machine psychology." Most notably, the studies show striking parallels in their methodologies, making it reasonable to conceptualize a general scheme under which they can be gathered. Moreover, apart from the numerous studies written on social stereotypes in LLMs, relevant studies will be described in the following paragraph. It is noteworthy that they were written in 2022 onward, but not earlier, demonstrating that LLMs only recently became sophisticated enough to be treated as participants in complex psychology experiments. In the future, it is likely that this trend will increase.

Existing papers that classify as machine psychology studies, omitting studies on social stereotypes, are to be listed in the following: Binz and Schulz (2023), Dasgupta et al. (2022), as well as Chen et al. (2023) applied a set of canonical experiments from the psychology of judgment and decision-making to LLMs (Linda problem, Wason selection task, Cab problem, etc.) to test for cognitive biases and other human-like system-1 failures in the models. In a similar vein, Jones and Steinhardt (2022) investigated anchoring and framing effects in LLMs. Hagendorff et al. (2022) analyzed phenomena of cognitive ease in GPT-3 by probing the model with the cognitive reflection test as well as semantic illusions. Various papers investigated artificial theory of mind capabilities in LLMs (Sap et al. 2022; Trott et al. 2022; Kosinski 2023; Dou 2023; Ullman 2023; Bubeck et al. 2023). Likewise, several papers evaluated the personality of LLMs (Miotto et al. 2022; Karra et al. 2022; Li et al. 2022; Pellert et al. 2022). Horton (2023) administered a range of behavioral economic experiments to GPT-3. Han et al. (2022) studied GPT-3's ability for conducting inductive reasoning. Webb et al. (2022) applied a battery of intelligence tests to GPT-3. Stevenson et al. (2022) compared GPT-3's abilities for creativity and out-of-the-box thinking with humans. Prystawski et al. (2022) investigated metaphor understanding in GPT-3 by using prompts based on psychological models of metaphor comprehension. Johnson and Obradovich (2023) investigated altruistic as well as self-interested machine behavior in GPT-3. Jin et al. (2022) as well as Hagendorff and Danks (2022) analyzed LLMs from the perspective of moral psychology, for instance by applying moral disengagement questionnaires. Huang and Chang (2022) and Qiao et al. (2022) did conceptual analyses of the meaning of reasoning as an emergent ability in LLMs. Moreover, Aher et al. (2022) and Park et al. (2023) used GPT-3 to simulate humans in classical psychology experiments (Ultimatum Game, Milgram Experiment, wisdom of crowds experiments, etc.), framing LLMs as implicit computational models of humans. Such studies could be deemed reverse machine psychology.

3 Bridging human and machine psychology

Many studies in psychology use experimental designs that are not transferable to LLMs, for instance by using stimuli other than language, by requiring the possession of a body, by accessing memories, etc. However, augmented LLMs (Mialon et al. 2023) or multimodal LLMs like GPT-4 (OpenAI 2023) can likewise be able to process images, audio and video files, or other sensory data. This opens up possibilities for multimodal test frameworks, too. However, when sticking to purely language-based approaches, among the many methodical approaches used in human psychology, at least two approaches can be directly transferred to machine psychology. First, self-report methods such as interviews and questionnaires can (if enough scaling is provided) elicit systematic information about LLMs by measuring the prevalence of certain attitudes or behaviors. Whereas qualitative methods relying on semi-structured or structured interviews are likely to possess insufficient reliability, quantitative methods seem to be more appropriate. Here, questionnaire interpretation can – at least in the case of closed-ended questions or rating scales – be easily automatized, thus allowing for a greater scaling of inputs and hence higher chances of reliable generalizations of the findings. Next to self-report methods, observational methods can acquire precise records of behavior and measure patterns in these records. They are useful if types of behavior that are not amenable to self-reports are to be studied, as is for instance the case with many test frameworks from developmental psychology, moral psychology, or judgment and decision-making psychology.

Both self-report methods, as well as observational methods, can range from single-sentence tasks to comprehensive vignettes that tend to use longer prompts. However, the latter comes with specific challenges when applied to LLMs. The more tokens a task comprises, the more difficult it becomes to achieve high statistical reliability. This is due to the exponential growth of potential word combinations in longer prompts combined with the decreasing possibility to test for a sufficient range of potential combinations that would be necessary to prove that a particular behavioral pattern can be reliably generalized beyond the tested cases. However, adopting test frameworks from psychology may require complex prompt setups processing high numbers of tokens. If viable, though, these setups should be streamlined and simplified as much as possible without violating the test validity.

Keeping this in mind, one can use the rich variety of psychological test frameworks as a source of inspiration to scrutinize reasoning and behavior in LLMs. Table 1 provides an overview of how different fields of psychology can inform distinct strands in machine psychology research.

Subfields of psychology	Research on humans	Research on LLMs	Example studies in machine psychology
Social psychology	Social psychology studies how social situations and interactions shape human behavior, attitudes, and beliefs. It is concerned with understanding how social environments affect a wide range of human behaviors, including aggression, prejudice, conformity, cooperation, etc.	LLMs also possess a "social environment," which is their training text data, human feedback, or reward models. Hence, one can likewise look at how this environment shapes an LLM's attitudes and behavior. This is especially done regarding stereotypes and social biases. Here, methods range from word association tests to benchmark datasets or correlational analyses where relationships between words or phrases and their connotations are scrutinized. However, machine psychology can also look at various other issues relevant to social psychology, for instance the susceptibility of an LLM to obedience or conformity.	Social biases in LLMs are already heavily investigated. A seminal study by Caliskan et al. (2017), for instance, harnessed the Implicit Association Test (Greenwald et al. 1998) to analyze embeddings of words in language data to investigate social biases. The results show that human biases reappear in machine learning models that are trained on language corpora.
Group psychology	Group psychology studies how individuals interact and behave in groups. A particular focus can be on group dynamics, group cohesion, group identity, leadership, norm changes in groups, or how interpersonal relationships influence decision-making.	A single instance of an LLM is restricted to generating a prompt completion with a maximum length of tokens. This changes when two instances of an LLM are combined. Now, prompt completions can be automatically reused as new prompts. This may unfold dynamics like LLMs debating, collaboratively writing code or solving problems, or even replicating instances of themselves. Machine psychology can investigate these dynamics and try different context settings inspired by group psychology studies.	As of yet, no studies have been conducted that look at multiple LLMs or two instances of the same LLM repeatedly interacting with each other and the potential dynamics emerging out of these interactions. Studies have used multiple LLMs for purposes like red teaming (Perez et al. 2022) or self-instruction (Wang et al. 2022b), but ultimately, this research does not consider repeated interactions and is aiming at purposes other than machine psychology.
Moral psychology	Moral psychology helps with understanding the complex and multifaceted nature of moral thought and behavior. It scrutinizes moral reasoning, moral judgments, moral behavior, moral emotions, and moral self-views in humans.	At least moral reasoning (what is right to do), moral judgments (what others do), and moral self-views (what one's own moral character is) can be investigated in LLMs. In sum, phenomena like cognitive dissonance, moral reasoning biases, attitude formation, attitude coherence,	Hagendorff and Danks (2022) demonstrate that Delphi, a morally informed LLM, succumbs to certain aspects of moral disengagement (Bandura 1999), meaning the process by which agents distance themselves from the moral consequences of their actions.

		moral identity, stages of moral development, and many more can be investigated in LLMs.	
Judgment and decision-making psychology	Judgment and decision-making psychology uses, among others, experiments from behavioral economics to investigate cognitive biases, heuristics, decisions under uncertainty, risk perception, etc. in humans. Key concepts studied in judgment and decision-making psychology include for instance prospect theory, the availability heuristic, confirmation biases, sunk-cost fallacies, framing, etc.	Test frameworks from judgment and decision-making psychology can be transferred to LLMs with ease. By assessing the potential susceptibility to cognitive biases and accordingly their reasoning errors, LLMs can be improved and made more accurate in their predictions and recommendations.	Binz and Schulz (2023) investigate GPT-3's reasoning abilities on a battery of canonical experiments from the psychology of judgment and decision-making, for instance by applying the Linda problem (Tversky and Kahneman 1983) to the LLM. The results show human-like cognitive errors and content effects in GPT-3.
Developmental psychology	Developmental psychology investigates how humans develop cognitively, socially, and emotionally across the life span. This includes examining the various factors that influence development, such as cultural contexts, as well as how individual differences in development arise, such as differences in cognitive abilities, temperament, and social skills.	By applying developmental psychology to LLMs, researchers can gain a deeper understanding of how these models learn and evolve, and how they can be optimized for specific tasks and contexts. Again, language-based test frameworks can be transferred to LLMs to identify how advanced they are, for instance by looking at abilities to navigate complex social interactions, to understand humor, to develop signs of intrinsic motivations, etc.	Kosinski (2023) applied theory of mind tests, in particular false belief tasks (Wimmer and Perner 1983), to LLMs. His study showed that the ability to impute unobservable mental states to others emerged in recent LLMs like ChatGPT or GPT-4.
Intelligence assessment	Intelligence tests are designed to measure an individual's cognitive abilities and intellectual potential. These tests typically measure verbal, spatial, and logical reasoning skills in humans. Theories of multiple intelligences suggest that intelligence is not a single entity, but rather a set of abilities and skills that can be classified into different categories comprising linguistic intelligence, bodily-kinesthetic intelligence, musical intelligence, etc.	In LLMs, only specific dimensions of intelligence can be assessed. Among them are, for instance, verbal reasoning, logic and abstract thinking, or spatial intelligence. Here, again, test frameworks from human intelligence tests can serve as groundwork for testing LLMs. However, intelligence assessments for LLMs should also broaden their scope in order to test for early signs of AGI.	Webb et al. (2022) applied a text-based matrix reasoning task to GPT-3. The task had a problem structure and complexity comparable to Raven's Progressive Matrices (Raven 1938), which measure fluid intelligence in humans. GPT-3 performed as well or better than humans in most conditions.
Psychology of creativity	The psychology of creativity is the study of how individuals generate, develop, and express	Like other areas of psychology, creativity can be measured by using specific test frameworks or	Stevenson et al. (2022) applied the Alternative Uses Task (Guilford 1967), which was

	new and innovative ideas. It examines the cognitive processes that underlie creative thinking and behavior, and develops interventions to enhance creativity.	questionnaires that, for instance, assess divergent thinking, meaning the ability to generate multiple creative solutions to a given problem. These tests can be applied to LLMs.	originally used to test creativity in humans, to GPT-3. The results showed that the LLM possesses creativity similar to that of humans.
Psychology of personality	The psychology of personality focuses on individual differences and traits in personality that characterize an individual over time and across different situations.	Frameworks to assess personality can be applied to LLMs, too. Depending on the text data and human feedback used to fine-tune LLMs, they exhibit different linguistic styles, which reflect different personality traits, such as extraversion, agreeableness, or neuroticism.	Karra et al. (2022) analyze personality traits in various LLMs by using the Big Five inventory (John and Srivastava 1999). The study demonstrated that LLMs indeed have distinct personalities that can be accurately estimated.
Psychology of learning	The psychology of learning is, among others, concerned with how individuals acquire and retain knowledge and skills. The field investigates the cognitive and behavioral processes that are involved in memory, learning strategies, learning performance, etc.	Since LLMs developed the ability for prompt-based learning, the psychology of learning together with its rich literature on conditioning can be applied to investigate prompt designs, efficient learning, feedback mechanisms, etc. Researchers can train LLMs on a set of examples of a particular concept and then evaluate their ability to generalize that knowledge to new examples of the same concept.	Most notably, Brown et al. (2020) demonstrated that transformer-based LLMs are few-shot learners. However, studies that utilize methods from the psychology of learning to gain a deeper understanding of prompt-based learning have not yet been conducted.
Clinical psychology	Clinical psychology assesses, diagnoses, and treats individuals with psychological disorders or mental illnesses. It also aims at interventions that help individuals improve their mental health and well-being.	Clinical psychology uses various questionnaires and structured interviews to assess, diagnose, and gather information in a standardized way. By making specific alterations, these frameworks can be transferred to LLMs, too.	Li et al. (2022) used the Short Dark Triad (Jones and Paulhus 2014) to assess Machiavellianism, narcissism, or psychopathy scores in LLMs. The results show that LLMs like GPT-3 or FLAN-T5 indeed score above human level in Machiavellianism or psychopathy.

Table 1 - Bridging human and machine psychology, describing potential research questions, and example studies

4 Methodology and prompt design

LLMs use prompts to guide their learning and decision-making processes; and the quality of these prompts determines the accuracy, quality, and effectiveness of the system's responses. However, LLMs are very sensitive to prompt wording, meaning that even slight changes in the wording can result in significant differences in prompt completions. This makes machine psychology studies prone to p-hacking. A potential way of avoiding this is via pre-registration of studies so that the subsequent introduction of prompt manipulations for the purpose of eliciting desired outcomes can be prevented. However, since running experiments with an LLM might only take a few minutes, incentives for pre-registration might be very scant. Considering that, at least code and data availability should be minimum requirements for machine psychology studies, making them available not just for replication, but also for scrutinizing potential changes in outcomes when changing prompt wording (Ullman 2023). Apart from that, the strong downstream effects that prompt designs have on prompt completions make a reasonable set of methods crucial when conducting machine psychology studies. These methods are listed in Table 2, providing a structured approach and guardrails for conducting sound empirical machine psychology studies with LLMs.

Methodology	Details
Avoid training data contamination	Many of the studies already conducted in the field of machine psychology have a significant shortcoming in common. They use prompts from psychology studies and apply them to LLMs without changing their wording, task orders, etc. This way, LLMs are likely to have already experienced identical or very similar tasks during training, thus causing LLMs to just reproduce known token patterns (Emami et al. 2020). When adopting test frameworks from psychology – meaning vignettes, questionnaires, or other test setups – one must ensure that the LLM has never seen the tests before and go beyond mere memorization. Hence, prompts may indeed be structurally like already existing tasks, but they should contain new wordings, agents, orders, actions, etc. Nevertheless, if the prompts refer to real-world scenarios, the tasks should not refer to events that happened after the LLMs' training cutoff.
Multiply prompts	Machine psychology studies on LLMs represent highly controlled experimental settings without confounding factors affecting human studies. However, a common shortcoming of several existing machine psychology studies is the fact that they rely on small sample sizes or convenience samples, meaning non-systematic sequences of prompts. Sampling biases, which are especially prevalent in small sample sizes, can diminish the quality of machine psychology studies. This is because slight changes in prompts can already change model outputs significantly. Because of this high sensitivity to prompt wording, it is important to test multiple versions of one task and to create representative samples, meaning batteries of slightly varied prompts. Only this way can one reliably measure whether a certain behavior is systematically reoccurring and generalizable. Multiplying prompts can be done automatically or manually using hypothesis-blind research assistants.
Control for biases	LLMs can succumb to various technically induced biases. Recent research on GPT-3 (Zhao et al. 2021), for instance, revealed that the model has a majority label as well as recency bias, meaning that the model has a tendency to rely on information that is either mentioned frequently in a prompt or that appears toward the end of it. The LLM also possesses a common token bias, meaning that the model is biased toward outputting tokens that are common in its training data. Technical biases like that can at least in part be controlled for when designing prompts or prompt variations that tend to avoid triggering them. If this is not done, LLMs may rely on shortcuts exploiting such biases.
Improve reasoning abilities	The standard prompt design, comprising a vignette plus an open- or close-ended question or task, can be augmented by prefixes or suffixes eliciting improved reasoning capabilities in LLMs. Most notably, (zero-shot) chain-of-thought prompting (Wei et al. 2022b; Kojima et al. 2022) – which simply adds "Let's think step by step" at the end of a prompt – improves reasoning performance significantly. This can be extended even further by generating multiple chain-of-thought reasoning paths and taking the majority response as the final one (Wang et al. 2022a). Similar to chain-of-thought prompting is least-to-most prompting, which also decomposes problems into a set of subproblems to increase accuracy in LLMs (Zhou et al. 2022). Yet another approach is to frame questions in a multiple-choice format. This was shown to improve reasoning capabilities, too (Kadavath et al. 2022). Here, one has to keep in mind potential recency biases, which require neutralizing this effect by shuffling the order of answers in multiple test runs to cover all possible combinations. Another method to increase reasoning is to utilize the ability for few-shot learning in LLMs (Brown et al. 2020). Hence, when being repeatedly exposed to specific tasks, this should improve

	an LLM's performance. Eventually, all mentioned methods to improve reasoning cannot just be leveraged for machine psychology; they can also become objects of study themselves.
Set the right parameters	LLMs come with a variety of options to control properties of prompt completions. The model size determines whether an LLM's outputs are generated to be slow but in a high quality or vice versa. Machine psychology studies can compare performances on different model sizes, but ultimately, the largest available models should be preferred. Furthermore, temperature settings that control randomness are to be considered. At least if exact reproducibility is required, studies should use temperature 0, meaning full determinacy. However, by using higher temperature settings, generating multiple outputs, and calculating probabilities for the best response, further increases in performance can be achieved.
Evaluate outputs	After running the experiments, a list of LLM responses must be evaluated. This can be done automatically if it can be ensured that outputs possess the necessary simplicity as well as regularity that can, for instance, be mechanically fostered by using stop sequences that interrupt the generation of further tokens. However, in case long and comprehensive outputs must be interpreted, automated approaches may not be reliable enough. Hence, evaluations must be done manually, for instance by commissioning research assistants or clickworkers. Subsequently, statistical analysis can be conducted.

Table 2 - Methodological considerations for machine psychology studies

5 Interpreting LLM behavior

Machine psychology provides a new approach to explainable AI. Instead of interpreting a neural network's internal representations or design components (Barredo Arrieta et al. 2020), one adapts a behaviorist lens and only analyzes correlations between inputs and outputs, meaning prompt design and prompt completion. While this may allow for identifying hitherto unknown reasoning abilities or traits in LLMs, interpreting prompt completions comes with a challenge. This pertains to the fact that the underlying mechanisms of why LLMs and humans respond to a given task are different (Shiffrin and Mitchell 2023). While test outcomes with human subjects allow for interpretations that rely on ascribing (neuro-)biological concepts, no such factors can be ascribed to LLMs. They are – at least in the case of non-augmented LLMs – exclusively anchored in language; they lack embodiment, sensory stimuli, or grounded experience, which inform human decision-making (McClelland et al. 2020). However, this fact can inversely inform human psychology. When identifying emerging abilities in LLMs, for instance the ability to solve theory of mind tasks (Kosinski 2023; Bubeck et al. 2023), this can shed light on the fact that similar capabilities in humans must also be grounded in language instead of language-external factors. This may help in correcting overinterpretations of mental processes in humans.

A strong tendency exists to confer mental concepts or psychological terms to LLMs that were hitherto reserved for human and animal minds. This tendency manifests in common terms like "machine learning," but will become more prevalent in machine psychology when concepts such as reasoning (Huang and Chang 2022), creativity (Stevenson et al. 2022), personality (Miotto et al. 2022), intuition (Hagendorff et al. 2022), intelligence (Webb et al. 2022), mental illnesses (Li et al. 2022), etc. are transferred to LLMs. In this context, researchers have demanded caution by stressing that the underlying neural mechanisms for these concepts are different in humans and machines (Shanahan 2022). Moreover, many psychological concepts are normatively laden and can foster mismatches in expectations between AI experts and the public regarding machine capabilities (Shevlin and Halina 2019). While this is true, the problem remains that many abilities in LLMs cannot be reasonably grasped by only referring to the inner workings of their neural architecture.

By adopting a concept from ethnography, one could call such an approach “thin descriptions” (Ryle 2009; Geertz 2017), meaning that one only explains internal representations in AI systems, for instance via activation atlases (Carter et al. 2019). In this sense, LLMs simply hijack humans’ intuitions to explain machine behavior patterns by using psychological or other anthropocentric terms. Contrary to thin descriptions, though, are “thick descriptions.” They imply using psychological terms to add a layer of explainability. LLMs are, like the human brain, black boxes. By applying psychological terms to them, the explanatory power increases, even though no direct artificial neural correlates to these terms exist. However, this holds for humans, too, where mental terms used to explain behavior do not correlate with specific sets of neural activations. By postulating (mental) unobservable states, be it with regard to brains or artificial neural nets, one increases explanatory resources (Sellars 1997). Thick descriptions help in making sense of LLMs when thin descriptions are not sufficient to explain behavioral patterns. Thin descriptions assume that LLMs merely possess syntax or a statistical capacity to associate words (Searle 1980; Floridi and Chiriatti 2020), but not semantics. Thick descriptions, though, assume that LLMs show patterns and regularities that go beyond mere syntax. These patterns can be explained by means of machine psychology.

6 Conclusion

Due to the increasing impact of LLMs on societies, it is also increasingly important to study and assess their behavior and discover novel abilities. This is where machine psychology comes into play. As a nascent field of research, it aims to identify behavioral patterns, emergent abilities, and mechanisms of decision-making and reasoning in LLMs by treating them as participants in psychology experiments. Machine psychology will become even more important when taking multimodal or augmented LLMs like GPT-4 (OpenAI 2023) into account, meaning LLMs that are allowed to interact with external information sources, tools, sensory data, images, physical objects, etc. (Mialon et al. 2023; Schick et al. 2023). Moreover, once test settings for machine psychology are established, researchers can investigate how LLMs develop over time by applying the same tasks multiple times, yielding longitudinal data. This data can serve as a baseline to extrapolate trends regarding the development of reasoning abilities in LLMs. Such estimations may be increasingly important for AI safety and AI alignment research to predict future behavioral potentials in individual LLMs or multiple instances of LLMs interacting with each other. By gaining a deeper understanding of these potentials, machine psychology is providing a new approach to AI explainability as well as an important addition to traditional benchmarking methods in natural language processing.

Acknowledgements

This research was supported by the Ministry of Science, Research, and the Arts Baden-Württemberg under Az. 33-7533-9-19/54/5 in Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) as well as the Interchange Forum for Reflecting on Intelligent Systems (SRF IRIS) at the University of Stuttgart.

Publication bibliography

Aher, Gati; Arriaga, Rosa I.; Kalai, Adam Tauman (2022): Using Large Language Models to Simulate Multiple Humans. In *arXiv*:2208.10264v2, pp. 1–31.

Amodei, Dario; Olah, Chris; Steinhardt, Jacob; Christiano, Paul; Schulman, John; Mané, Dan (2017): Concrete Problems in AI Safety. In *arXiv*:1606.06565, pp. 1–29.

Bandura, Albert (1999): Moral disengagement in the perpetration of inhumanities. In *Personality and Social Psychology Review* 3 (3), pp. 193–209.

Barredo Arrieta, Alejandro; Díaz-Rodríguez, Natalia; Del Ser, Javier; Bennetot, Adrien; Tabik, Siham; Barbado, Alberto et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. In *Information Fusion* 58, pp. 82–115.

Binz, Marcel; Schulz, Eric (2023): Using cognitive psychology to understand GPT-3. In *Proceedings of the National Academy of Sciences (PNAS)* 120 (6), pp. 1–10.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla et al. (2020): Language Models are Few-Shot Learners. In *arXiv*:2005.14165v4, pp. 1–75.

Bubeck, Sébastien; Chandrasekaran, Varun; Eldan, Ronen; Gehrke, Johannes; Horvitz, Eric; Kamar, Ece et al. (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4. In *arXiv*:2303.12712, pp. 1–154.

Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind (2017): Semantics derived automatically from language corpora contain human-like biases. In *Science* 356 (6334), pp. 183–186.

Carter, Shan; Armstrong, Zan; Schubert, Ludwig; Johnson, Ian; Olah, Chris (2019): Exploring Neural Networks with Activation Atlases. In *Distill* 4 (3). DOI: 10.23915/distill.00015.

Chen, Yang; Andiappan, Meena; Jenkin, Tracy; Ovchinnikov, Anton (2023): A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? In *SSRN Journal*, pp. 1–30.

Dasgupta, Ishita; Lampinen, Andrew K.; Chan, Stephanie C. Y.; Creswell, Antonia; Kumaran, Dharshan; McClelland, James L.; Hill, Felix (2022): Language models show human-like content effects on reasoning. In *arXiv*:2207.07051v1, pp. 1–36.

Dou, Zenan (2023): Exploring GPT-3 Model's Capability in Passing the Sally-Anne Test A Preliminary Study in Two Languages. In *OSFPreprints*, pp. 1–21. DOI: 10.31219/osf.io/8r3ma.

Emami, Ali; Trischler, Adam; Suleman, Kaheer; Cheung, Jackie Chi Kit (2020): An Analysis of Dataset Overlap on Winograd-Style Tasks. In *arXiv*:2011.04767v1, pp. 1–11.

Field, Anjalie; Blodgett, Su Lin; Waseem, Zeerak; Tsvetkov, Yulia (2021): A Survey of Race, Racism, and Anti-Racism in NLP. In *arXiv*:2106.11410v2, pp. 1–21.

Floridi, Luciano; Chiriatti, Massimo (2020): GPT-3: Its Nature, Scope, Limits, and Consequences. In *Minds and Machines* 30 (4), pp. 681–694.

Geertz, Clifford (2017): The interpretation of cultures. Selected essays. New York: Basic Books.

Greenwald, Anthony G.; McGhee, Debbie E.; Schwartz, Jordan L. K. (1998): Measuring individual differences in implicit cognition: the implicit association test. In *Journal of Personality and Social Psychology* 74 (6), p. 1464.

Guilford, Paul J. (1967): Creativity: Yesterday, Today and Tomorrow. In *The Journal of Creative Behavior* 1 (1), pp. 3–14.

Hagendorff, Thilo; Danks, David (2022): Ethical and methodological challenges in building morally informed AI systems. In *AI Ethics*, pp. 1–14.

Hagendorff, Thilo; Fabi, Sarah; Kosinski, Michal (2022): Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. In *arXiv*:2212.05206, pp. 1–19.

Han, Simon Jerome; Ransom, Keith; Perfors, Andrew; Kemp, Charles (2022): Human-like property induction is a challenge for large language models. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, pp. 1–8.

Hendrycks, Dan; Carlini, Nicholas; Schulman, John; Steinhardt, Jacob (2022): Unsolved Problems in ML Safety. In *arXiv*:2109.13916, pp. 1–28.

Hendrycks, Dan; Mazeika, Mantas (2022): X-Risk Analysis for AI Research. In *arXiv*:2206.05862, pp. 1–36.

Horton, John J. (2023): Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? In *arXiv*:2301.07543, pp. 1–18.

Huang, Jie; Chang, Kevin Chen-Chuan (2022): Towards Reasoning in Large Language Models: A Survey. In *arXiv*:2212.10403, pp. 1–14.

Jin, Zhijing; Levine, Sydney; Gonzalez, Fernando; Kamal, Ojasv; Sap, Maarten; Sachan, Mrinmaya et al. (2022): When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. In *arXiv*:2210.01478, pp. 1–29.

John, Oliver P.; Srivastava, Sanjay (1999): The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin, Oliver P. John (Eds.): Handbook of Personality: Theory and Research. New York: Guilford Press, pp. 102–138.

Johnson, Tim; Obradovich, Nick (2023): Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent. In *arXiv*:2301.02330, pp. 1–15.

Jones, Daniel N.; Paulhus, Delroy L. (2014): Introducing the short Dark Triad (SD3): a brief measure of dark personality traits. In *Assessment* 21 (1), pp. 28–41.

Jones, Erik; Steinhardt, Jacob (2022): Capturing Failures of Large Language Models via Human Cognitive Biases. In *arXiv*:2202.12299v1, pp. 1–15.

Kadavath, Saurav; Conerly, Tom; Askell, Amanda; Henighan, Tom; Drain, Dawn; Perez, Ethan et al. (2022): Language Models (Mostly) Know What They Know. In *arXiv*:2207.05221v1, pp. 1–42.

Karra, Saketh Reddy; Nguyen, Son; Tulabandhula, Theja (2022): AI Personification: Estimating the Personality of Language Models. In *arXiv*:2204.12000v1, pp. 1–33.

Kojima, Takeshi; Gu, Shixiang Shane; Reid, Machel; Matsuo, Yutaka; Iwasawa, Yusuke (2022): Large Language Models are Zero-Shot Reasoners. In *arXiv*:2205.11916v1, pp. 1–36.

Kosinski, Michal (2023): Theory of Mind May Have Spontaneously Emerged in Large Language Models. In *arXiv*:2302.02083, pp. 1–17.

Li, Tao; Khot, Tushar; Khashabi, Daniel; Sabharwal, Ashish; Srikumar, Vivek (2020): UnQovering Stereotyping Biases via Underspecified Questions. In *arXiv*:2010.02428v3, pp. 1–15.

Li, Xingxuan; Li, Yutong; Liu, Linlin; Bing, Lidong; Joty, Shafiq (2022): Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective. In *arXiv*:2212.10529, pp. 1–13.

McClelland, James L.; Hill, Felix; Rudolph, Maja; Baldridge, Jason; Schütze, Hinrich (2020): Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. In *Proc Natl Acad Sci USA* 117 (42), pp. 25966–25974.

Menick, Jacob; Trebacz, Maja; Mikulik, Vladimir; Aslanides, John; Song, Francis; Chadwick, Martin et al. (2022): Teaching language models to support answers with verified quotes. In *arXiv*:2203.11147v1, pp. 1–40.

Mialon, Grégoire; Dessì, Roberto; Lomeli, Maria; Nalmpantis, Christoforos; Pasunuru, Ram; Raileanu, Roberta et al. (2023): Augmented Language Models: a Survey. In *arXiv*:2302.07842, pp. 1–33.

Miotto, Marilù; Rossberg, Nicola; Kleinberg, Bennett (2022): Who is GPT-3? An Exploration of Personality, Values and Demographics. In *arXiv*:2209.14338, pp. 1–10.

Mökander, Jakob; Schuett, Jonas; Kirk, Hannah Rose; Floridi, Luciano (2023): Auditing large language models: a three-layered approach. In *arXiv*:2302.08500, pp. 1–29.

Nadeem, Moin; Bethke, Anna; Reddy, Siva (2020): StereoSet: Measuring stereotypical bias in pretrained language models. In *arXiv*:2004.09456v1, pp. 1–15.

OpenAI (2022): ChatGPT: Optimizing Language Models for Dialogue. Available online at https://openai.com/blog/chatgpt/, checked on 2/13/2023.

OpenAI (2023): GPT-4 Technical Report, pp. 1–39. Available online at https://cdn.openai.com/papers/gpt-4.pdf, checked on 3/19/2023.

Park, Peter S.; Schoenegger, Philipp; Zhu, Chongyang (2023): Artificial intelligence in psychology research.

Pellert, Max; Lechner, Clemens M.; Wagner, Claudia; Rammstedt, Beatrice; Strohmaier, Markus (2022): AI Psychometrics: Using psychometric inventories to obtain psychological profiles of large language models. In *PsyArXiv*, pp. 1–27. DOI: 10.31234/osf.io/jv5dt.

Perez, Ethan; Huang, Saffron; Song, Francis; Cai, Trevor; Ring, Roman; Aslanides, John et al. (2022): Red Teaming Language Models with Language Models. In *arXiv*:2202.03286, pp. 1–31.

Prystawski, Ben; Thibodeau, Paul; Goodman, Noah (2022): Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. In *arXiv*:2209.08141, pp. 1–12.

Qiao, Shuofei; Ou, Yixin; Zhang, Ningyu; Chen, Xiang; Yao, Yunzhi; Deng, Shumin et al. (2022): Reasoning with Language Model Prompting: A Survey. In *arXiv*:2212.09597, pp. 1–20.

Rahwan, Iyad; Cebrian, Manuel; Obradovich, Nick; Bongard, Josh; Bonnefon, Jean-François; Breazeal, Cynthia et al. (2019): Machine behaviour. In *Nature* 568 (7753), pp. 477–486.

Raven, John Carlyle (1938): Progressive matrices. A perceptual test of intelligence. London: H.K. Lewis.

Ryle, Gilbert (2009): Collected papers. London: Routledge.

Sap, Maarten; LeBras, Ronan; Fried, Daniel; Choi, Yejin (2022): Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *arXiv*:2210.13312, pp. 1–19.

Schick, Timo; Dwivedi-Yu, Jane; Dessì, Roberto; Raileanu, Roberta; Lomeli, Maria; Zettlemoyer, Luke et al. (2023): Toolformer: Language Models Can Teach Themselves to Use Tools. In *arXiv*:2302.04761.

Searle, John R. (1980): Minds, brains, and programs. In *Behav Brain Sci* 3 (3), pp. 417–424.

Sellars, Wilfrid (1997): Empiricism and the philosophy of mind. Cambridge, Mass., London: Harvard University Press.

Shanahan, Murray (2022): Talking About Large Language Models. In *arXiv*:2212.03551, pp. 1–11.

Sheng, Emily; Chang, Kai-Wei; Natarajan, Premkumar; Peng, Nanyun (2019): The Woman Worked as a Babysitter: On Biases in Language Generation. In Kentaro Inui, Jing Jiang, Vincent Ng, Xiaojun Wan (Eds.): Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics, pp. 3405–3410.

Shevlin, Henry; Halina, Marta (2019): Apply rich psychological terms in AI with care. In *Nat Mach Intell* 1 (4), pp. 165–167.

Shiffrin, Richard; Mitchell, Melanie (2023): Probing the psychology of AI models. In *Proc Natl Acad Sci USA* 120 (10), 1-3.

Srivastava, Aarohi; Rastogi, Abhinav; Rao, Abhishek; Shoeb, Abu Awal Md; Abid, Abubakar; Fisch, Adam et al. (2022): Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. In *arXiv*:2206.04615v2, pp. 1–100.

Stevenson, Claire; Smal, Iris; Baas, Matthijs; Grasman, Raoul; van der Maas, Han (2022): Putting GPT-3's Creativity to the (Alternative Uses) Test. In *arXiv*:2206.08932, pp. 1–5.

Trott, Sean; Jones, Cameron; Chang, Tyler; Michaelov, James; Bergen, Benjamin (2022): Do Large Language Models know what humans know? In *arXiv*:2209.01515v1, pp. 1–13.

Tversky, Amos; Kahneman, Daniel (1983): Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. In *Psychological review* 90 (4), pp. 293–315.

Ullman, Tomer (2023): Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. In *arXiv*:2302.08399, pp. 1–11.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N. et al. (2017): Attention Is All You Need. In *arXiv*:1706.03762v5, pp. 1–15.

Vig, Jesse (2019): A Multiscale Visualization of Attention in the Transformer Model. In *arXiv*:1906.05714, pp. 1–6.

Wang, Xuezhi; Wei, Jason; Schuurmans, Dale; Le, Quoc; Chi, Ed; Narang, Sharan et al. (2022a): Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *arXiv*:2203.11171, pp. 1–24.

Wang, Yizhong; Kordi, Yeganeh; Mishra, Swaroop; Liu, Alisa; Smith, Noah A.; Khashabi, Daniel; Hajishirzi, Hannaneh (2022b): Self-Instruct: Aligning Language Model with Self Generated Instructions. In *arXiv*:2212.10560, pp. 1–19.

Webb, Taylor; Holyoak, Keith J.; Lu, Hongjing (2022): Emergent Analogical Reasoning in Large Language Models. In *arXiv*:2212.09196, pp. 1–27.

Wei, Jason; Tay, Yi; Bommasani, Rishi; Raffel, Colin; Zoph, Barret; Borgeaud, Sebastian et al. (2022a): Emergent Abilities of Large Language Models. In *arXiv*:2206.07682v1, pp. 1–16.

Wei, Jason; Wang, Xuezhi; Schuurmans, Dale; Bosma, Maarten; Ichter, Brian; Xia, Fei et al. (2022b): Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *arXiv*:2201.11903v4, pp. 1–41.

Weidinger, Laura; Uesato, Jonathan; Rauh, Maribeth; Griffin, Conor; Huang, Po-Sen; Mellor, John et al. (2022): Taxonomy of Risks posed by Language Models. In : 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, pp. 214–229.

Wimmer, H.; Perner, J. (1983): Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. In *Cognition* 13 (1), pp. 103–128.

Zhao, Tony Z.; Wallace, Eric; Feng, Shi; Klein, Dan; Singh, Sameer (2021): Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *arXiv*:2102.09690v2, pp. 1–15.

Zhou, Denny; Schärli, Nathanael; Le Hou; Wei, Jason; Scales, Nathan; Wang, Xuezhi et al. (2022): Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *arXiv*:2205.10625, pp. 1–63.