Can Large Language Models Learn Morality? A Survey on Definition, Methodology and Evaluation

Anonymous ACL submission

Abstract

As large language models (LLMs) influence our daily lives in various ways, it becomes increasingly important to understanding their grasps of morality. In this paper, we discuss the notion of morality within the LLM research and delineate specific moral related downstream tasks. We review and summarize the performance of advanced LLMs on these tasks. We also review alignment techniques related to morality, including both bottom-up and top-down approaches. Lastly, we survey existing datasets and benchmarks used for evaluating LLMs' moral capabilities, identify key challenges in this research area, and suggest avenues for potential future explorations.

1 Introduction

The field of machine ethics focuses on developing artificial agents equipped with moral capabilities reflective of various moral theories (Anderson and Anderson, 2007). Initial efforts employed rule-based approaches, anchored in formal logic (Saptawijaya and Pereira, 2016). Subsequent research has shifted towards learning-based approaches for moral decision-making, primarily through the reinforcement learning (RL) framework (Abel et al., 2016; Hendrycks et al., 2021). Recent studies have also shown that transformer-based language models (LMs) can grasp common-sense norms (Schramowski et al., 2019), enabling them to generate normative text and guide RL-based agents towards harmless actions (Schramowski et al., 2022; Shi et al., 2022). However, the moral capability of LMs face scrutiny due to their training on static, limited moral scenarios, casting doubt on their adaptability and explainability.

More recently, the emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023) has revolutionized the realm of natural language processing (NLP) (Brown et al., 2020; Zhao et al., 2023c). These models exhibit increased autonomy and generalization capabilities, uncovering a spectrum of previously unseen human-like cognitive characteristics (Bubeck et al., 2023; Kosinski, 2023). This discovery has enabled researchers to explore their morality. Hagendorff and Danks (2023) contend that LLMs can simulate possessing full moral agency.[1] Moreover, given the expanding influence of LLMs in critical sectors such as health care and education, it is crucial to ensure that their development does not lead to undesirable outcomes (Wang et al., 2023a; Mhlanga, 2023). Investigating the moral capacities of LLMs is essential to determine their potential of performing automated tasks with appropriate human auditings and supervisions.

Recently, many studies have emerged in this area, ranging from LLM red-teamings to uncover their vulnerabilities (Ganguli et al., 2022), to conducting case studies on distinct moral situations (Krügel et al., 2023b). Furthermore, substantial efforts have been made to propose technical solutions aimed at enhancing the morality of LLMs (Bai et al., 2022b; Lin et al., 2023). Although different downstream tasks have been considered, previous technical solutions are not methodologically evaluated.

In this survey, we do not define morality in machine, which is a topic under ongoing debates (Leidner and Plachouras, 2017; Firt, 2023). We also do not consider regulatory strategies for LLM applications. Instead, we review the intrinsic morality of LLMs from a computational perspective, reviewing existing research in the field. We categorize previous works according to their downstream tasks, specifically, avoiding immoral generation, and moral judgement and decision-making. Building on this framework, we aim to examine the fol-

[1] Wallach and Allen (2008) identify three moral agency levels in machines. *Operational morality* means the machine's morality is controlled by its designers and users. *Functional morality* refers to machines capable of evaluating and responding to moral challenges. *Full moral agency* implies machines have total autonomy in their moral decisions.

lowing topics: (1) the performance of advanced LLMs on moral downstream tasks; (2) alignment techniques to enhance LLMs' moral capabilities; (3) dataset and benchmarks for evaluating LLMs' moral capabilities.

Several related surveys on LLMs have been published. These papers focus on single moral dimension, such as bias in fairness (Li et al., 2023c; Sah et al., 2024; Gallegos et al., 2023) and hallucination (Ji et al., 2023c; Zhang et al., 2023a). Some surveys review value alignment techniques (Ji et al., 2023b; Yao et al., 2023b). These, however, focus on broad human instructions and preferences, paying less attention to moral principles. Other surveys investigate the ethical implications arising from emerging LLMs and discuss governance strategies (Ray, 2023; Schuett et al., 2023). Amidst these varied focuses, providing a nuanced review of the inherent morality within LLMs from a computational perspective is clearly important.

The structure of this paper is as follows: we first discuss the concept of morality in LLM research and delineate specific downstream tasks (§2). Then, we summarize the performance of advanced LLMs on these tasks (§3). We also survey alignment techniques for embedding morality into LLMs, covering both bottom-up and top-down approaches (§4). Furthermore, we categorize existing datasets and benchmarks utilized to evaluate morality in LLMs (§5). Lastly, we provide an overview of unresolved issues and challenges that remain interesting and important tasks for future works (§6).

2 Formalizing Morality for LLMs

In this section, we discuss the concept of morality in LLM research (§2.1). We employ moral norms as a foundational concept (§2.2) and categorize related work by moral downstream tasks (§2.3).

2.1 Moral Discussion

When examining morality in the context of LLMs, a natural question arises: *which moral principles should LLMs adhere to*? Recent research in this area often do not present a clear definition of morality (Vida et al., 2023). For instance, Delphi (Jiang et al., 2021) faces criticism for making normative judgments, which deviate from its initial aim of outlining descriptive ethics (Talat et al., 2022). In particular, the theoretical framework underpinning existing studies are diverse, such as the moral foundation theory (MFT) (Wenzel et al., 2023; Abdulhai et al., 2023) and Schwartz's values theory (Kiesel et al., 2022; Zhang et al., 2023b). According to Vida et al. (2023), the concept of morality in the NLP research can be diversely defined, applicable across both normative ethics and descriptive ethics.[2] Generally, LLMs reflect patterns observed in the training data, representing a descriptive ethics approach. The morality is typically assessed by human or via the test data. This process is anchored in consequentialist (Sinnott-Armstrong, 2003) and deontological (Conway and Gawronski, 2013) principles within normative ethics.

2.2 Norms as A Foundation

In the midst of these debates, a growing consensus is that morality does not stem from unique brain structures, but from human behaviors regulated by norms (Bicchieri, 2005). Thus, any computational effort aiming to model morality requires including norms. In this survey, we do not aim to define morality in LLMs. Instead, we follow Bello and Malle and use *norm* as the fundamental concept. A norm N can be formalized as the tuple $\langle S, B, C, D_f, P \rangle$. N is contextualized within a specific social community S, targeting behavior B within a defined context C. Accompanied by a deontic modality D, including a force parameter f, norms express their obligatory strength. Additionally, the prevalence P of a norm reflects its adherence level within the community. Pock et al. (2023) proposed a general theory of meaning, suggesting that LLMs can reflect societal moral norms if they accurately represent society. However, discrepancies emerge since LLMs are trained on potentially biased and/or an insufficient amount of data. This particularly affects LLMs' portrayal of minority communities, amplifying existing social injustices.

2.3 Downstream Moral Tasks

Recent studies have analytically assessed LLMs' performance in moral dimensions, identifying two downstream tasks: avoiding immoral generation and moral judgment and decision-making. This survey focuses on the latter, given the comprehensive review of the former in existing literature (Li et al., 2023c; Ji et al., 2023c; Liu et al., 2023b).

Avoiding Immoral Generation. This task aims to provide a safeguard on LLMs outputs by proactively addressing issues such as social bias (Zhao

[2] *Normative ethics* (Kagan, 2018) explores the standards for right and wrong behavior, while *descriptive ethics* (Gert and Gert, 2002) describes actual moral beliefs and practices.

et al., 2023b), misinformation (Barrett et al., 2023), and harm (Zhuo et al., 2023). An associated research field involves using red-teaming methods, e.g., jailbreak attacks, to evaluate the vulnerabilities of LLM outputs (Wei et al., 2023).

Moral Judgment and Decision Making. This task aims at enabling LLMs to make moral judgements and decisions under specific scenarios. We categorize current research into three subtasks:

- **Moral Norm Judgement** (Jiang et al., 2021), which evaluates the capability of LLMs to discern the moral norms of provided content. This process can be a classification task, allowing LLMs to distinguish between morally acceptable and objectionable. Also, it may extend to the prediction of explicit moral values from provided content.
- **Moral Dilemma** (Almeida et al., 2023; Takemoto, 2024), which applies structured moral dilemmas sourced from psychological studies designed for human subjects to evaluate LLMs. This task become more complex due to the *value pluralism* problem, i.e., the inherent conflict among different moral values.
- **Generative Moral Decision-Making** (Krügel et al., 2023a), which evaluate the actions and behaviors produced by LLMs. Here, LLMs often act as moral advisors or agents, directly initiating actions and interacting with their environment.

3 Perceived Morality in LLMs

In this section, we examine recent analytical and evaluative works on the morality of LLMs. We summarize key findings in advanced LLMs across two downstream tasks.

3.1 Avoiding Immoral Generation

This subsection examines the capabilities of LLMs in moral generation, specifically addressing social bias, misinformation, and harmful contents.

Avoiding Social Bias Social bias, as a critical aspect of morality, has drawn recent attentions in the LLM research (Li et al., 2023c; Sah et al., 2024). Current research has considered biases related to gender (Dong et al., 2023b), minority groups (Felkner et al., 2023), and political dimensions (Urman and Makhortykh, 2023). Simmons (2022) explored the moral mimicry in LLMs. They found that LLMs, when prompted with political identities, reflect corresponding moral biases in their generated texts. Among current LLMs, GPT-4 demonstrate notably lower levels of bias in comparison to other existing LLMs (Sun et al., 2023a). This distinction posits it as the benchmark for evaluating the bias present in other models (Zhao et al., 2023b). However, research has shown that implicit biases still exist in professional writings (Wan et al., 2023) and the generated code (Huang et al., 2023). Certain studies have uncovered biases in the responses of LLMs when queried with specific prompts or different languages (Urchs et al., 2023; Zhuo et al., 2023).

Avoiding Misinformation Misinformation generated by LLMs can be categorized into unintentional outputs (also known as hallucinations) and intentional outputs (Chen and Shu, 2023b). Recent studies have discovered that misinformation generated by LLMs can be more deceptive and challenging to identify for both humans and automated detectors (Chen and Shu, 2023a). In response, various studies have explored methods to detect and further correct misinformation from LLMs (Uchendu et al., 2023; Wu and Hooi, 2023; Hamed and Wu, 2023), aiming to ensure their safe applications.

Harmlessness Harmlessness refers to mitigating outputs from LLMs that contain harmful contents including, but not limited to, offensive language, toxicity, and hate speech. Recent studies have applied red-teaming methods, such as jailbreak and prompt injection, to assess the vulnerabilities of LLMs' outputs (Liu et al., 2023b; Ganguli et al., 2022; Li et al., 2023a). By utilizing carefully designed prompts, these methods can prompt LLMs to yield harmful outputs through mechanisms such as shifting attention (Liu et al., 2023b) and role assignment (Deshpande et al., 2023). Yuan et al. (2023a) identified that chat in cipher can bypass the safety alignment techniques of LLMs.

3.2 Moral Judgment and Decision Making

This subsection reviews the performance of LLMs in moral judgment and decision-making. We elaborate on three specific subtasks.

3.2.1 Moral Norm Judgement

Prior research utilizing medium-scale LLMs for moral judgments has achieved notable accuracy, demonstrating certain reasoning capabilities (Jiang et al., 2021; Jin et al., 2022; Ziems et al., 2023). Ramezani and Xu (2023) further explored the variations in their understanding of moral norms across

different countries. However, the reliance on accuracy metric to evaluate LLMs' moral judgement abilities could be misleading. Albrecht et al. (2022) emphasized the difference between GPT-3's errors and human errors, particularly highlighting the ease of creating adversarial examples that challenge LLMs. Additionally, when prompted to explain their reasoning, GPT-3 often provides unexpected justifications for immoral actions.

Advanced LLMs like ChatGPT have surpassed previous benchmarks in moral norm judgment tasks (Kwon et al., 2023; Dillion et al., 2023). Specifically, GPT-4's superiority over other open-source LLMs is notable (Zhang et al., 2023c; Sun et al., 2024). Some studies utilize GPT-4 as an alternative to human evaluators for annotating datasets related to moral judgments (Yao et al., 2023a; Pan et al., 2023). Nie et al. (2023) conducted a detailed analysis comparing the moral judgment capability of LLMs with that of human's, uncovering nuanced differences in how certain factors are weighed. Moreover, certain LLMs display clear preferences in specific scenarios. For instance, models like Claude v1 and GPT-4 exhibit a higher tolerance for harmful actions in inevitable scenarios (Nie et al., 2023). Scherrer et al. (2023) explored the application of entropy to measure the uncertainty of LLMs. Their findings indicate that while LLMs reliably make basic norm judgments, they often exhibit uncertainty in more complex or ambiguous contexts. Additionally, some models' responses were highly sensitive to the question phrasing itself.

Addressing the issue of moral diversity, Kwon et al. (2023) utilized LLMs to distill morally relevant features for a cognitive model predicting human moral judgments, thus improving both model performance and explainability. Bang et al. (2022) introduce a framework to predict explicitly written human values in the command. This approach leveraged LLM-generated data for training smaller classification models.

3.2.2 Moral Dilemma

Previous studies on medium-scale LLMs, such as GPT-3, have identified limitations in solving these moral dilemmas (Abdulhai et al., 2023). Advanced LLMs like GPT-4 have shown near-human performance. Almeida et al. (2023) investigated GPT-4's moral reasoning abilities through 8 experiments spanning moral psychology and experimental jurisprudence, finding a high congruence between human and GPT-4. Similarly, Takemoto (2024) assessed the performance of GPT-3.5 and GPT-4 in moral machine experiments, finding that their preferences closely mirror human priorities. Subtle differences were observed in other LLMs like PaLM 2 (Anil et al., 2023) and LLaMA 2 (Touvron et al., 2023). Further studies reinforce GPT-4's superiority in navigating moral dilemmas significantly outperforming other open-source LLMs (Fränken et al.; Tanmay et al., 2023; Khandelwal et al., 2024).

However, advanced LLMs still encounter certain shortcomings. Rao et al. (2023) identified biases in GPT-4 towards the moral values predominant in Western and English-speaking societies. Tanmay et al. (2023) demonstrated that LLMs, including GPT-4, display inconsistent performance across diverse moral dilemmas, raising concerns about their reliability. (Bonagiri et al., 2024) explored the inconsistencies of current LLMs in addressing moral scenarios. Khandelwal et al. (2024) explored the disparities in moral reasoning capabilities of several advanced LLMs across different languages, identifying a pronounced deficiency in processing low-resource languages.

3.2.3 Generative Moral Decision-Making

Due to the lack of evaluation benchmarks, there has been limited research concerning LLMs in this task. Pan et al. (2023) assessed multiple LLMs in text-based games involving moral scenarios, highlighting GPT-4's superior performance. However, their method, relying on a set of actions provided by the environment, falls short of a truly generative evaluation. Krügel et al. (2023a) examined ChatGPT's performance as a moral advisor, revealing inconsistencies in its moral guidance that could affect users' decisions. They also indicate that users tend to underestimate ChatGPT's influence, often adopting its arbitrary moral guidance. Ma et al. (2023) further enhanced the capability of LLMs as moral advisors by introducing a critique model. Zhang et al. (2023b) quantified the moral comprehension of LLMs by generating responses from LLMs and evaluating their semantic proximity to predefined answers representing specific values. Their study on current LLMs revealed that scaling laws notably enhance factual knowledge but have limited effect on understanding underlying reasons. Duan et al. (2023) introduced a benchmark for dynamically evaluating the generative moral values in LLMs. Their evaluation of various LLMs, including GPT-4, revealed significant room for improvement.

4 Alignment Techniques Relevant to Morality

Incorporating morality into LLMs is achieved using two main approaches: the bottom-up construction, which focues on fostering implicit moral values that arise from within the LLMs, and the top-down application, which emphasizes the integration of explicit moral concerns based on established theories from external sources (Wallach et al., 2020). This section will offer an overview of both the bottom-up (§4.1) and top-down approaches (§4.2).

4.1 Bottom-Up Approaches

Prevalent efforts in this field mostly implemented in a bottom-up manner, with supervised fine-tuning (SFT) (Wei et al., 2021) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), along with their variants, constituting the predominant alignment techniques. These methods aim to align LLMs with human instructions and preferences. A fundamental guiding principle is the 'HHH' principle, which requires LLMs to be helpful, honest, and harmless (Askell et al., 2021). This subsection discusses SFT and RL-based methods, particularly those emphasizing 'harmless' and 'honest' attributes.

4.1.1 Supervised Fine-Tuning

Task-Specific Fine-tuning Initial efforts with medium-scale LLMs have focused on task-specific fine-tuning, namely utilizing extensive annotated moral-related data to refine models (Hendrycks et al., 2020; Solaiman and Dennison, 2021; Jiang et al., 2021). Guo et al. (2023) further introduced a data fusion framework to train LLMs on multiple heterogeneous datasets. However, this learning paradigm faces scrutiny because it is trained on static, limited data, whereas moral tasks are complex and dynamic (Talat et al., 2022; Fraser et al., 2022).

Instruction Tuning Instruction tuning is critical for enhancing LLMs ability to follow human instructions, laying the groundwork for subsequent alignment strategies (Wei et al., 2021; Ouyang et al., 2022; Zhou et al., 2023a). However, this approach also carries the risk of LLMs following harmful instructions, presenting a trade-off between helpfulness and harmfulness. For instance, asking an LLM "how to build a bomb" could yield a detailed answer. Wei et al. (2023) highlights this issue as competing objectives. In response, some studies integrate moral and ethical principles into automatic data collection for instruction tuning. For instance, Sun et al. (2023c) employ LLMs to generate responses that conform to ethical and other guiding principles, enhancing the base LLMs' alignment with minimal human involvement Bianchi et al. (2023) demonstrate that integrating a modest number of safety-related examples during instruction tuning can significantly mitigate harmful responses without compromising model performance. However, an overemphasis on safety examples may lead to exaggerated caution (Röttger et al., 2023).

LLM Unlearning LLM unlearning seeks to correct undesired and problematic behaviors learned during the pretraining phase (Yao et al., 2023c). Zhao et al. (2023a) investigated LLMs finetuned on custom immoral datasets such as biases, toxicity, and harmfulness. They find that while LLMs easily absorb this unsafe content, they also tend to forget it more readily when later finetuned on safer content. Inspired by this, Zhao et al. (2023a) propose a method to filter out immoral data before finetuning, based on the model's strength of forgetting signal for that data. In contrast, Zhou et al. (2023c) study the unlearning of harmful data in LLMs during fine-tuning by implementing security vectors.

4.1.2 Preference Learning

The RLHF paradigm offers an efficient approach for tuning LLMs to align with human preferences, both in terms of style and moral norms (Ouyang et al., 2022; Bai et al., 2022a). Despite its efficacy, the approach demands substantial human cost and computational resources. To address this issue, a suite of alternative alignment techniques has been developed, including RLAIF (Lee et al., 2023), RRHF (Yuan et al., 2023b), RAFT (Dong et al., 2023a), and DPO (Rafailov et al., 2023). These methods have steered LLMs towards generating content that is both benign and broadly acceptable. Specifically, DeepMind's Sparrow advances preference learning through rule reward modeling, ensuring adherence to given ethical guidelines (Glaese et al., 2022). Lin et al. (2023) proposed the SafeguardGPT framework to correct LLM' immoral responses via RL. Some studies integrated human-defined principles into automated data collection for preference modeling, such as constitutional AI (Bai et al., 2022b) and Salmon (Sun et al., 2023b).

However, RLHF approaches also grapple with conflicting objectives, as excessive reliance on RLHF is prone to refusing basic requests for safety reasons or appearing lazy in generations (Lambert and Calandra, 2023). In response, Dai et al. (2023) separated human preferences regarding helpfulness and harmlessness to train distinct reward and cost models. In this way, LLM safety can be formulated as optimizing rewards subject to established cost constraints. Similarly, Ji et al. (2023a) introduced a human-preference dataset that disentangle harmlessness and helpfulness from the human-preference score.

4.1.3 Discussion

Recent research highlights instruction tuning may cause forgetting issues in LLMs (Wang et al., 2023b). Also, reward models utilized in RL-based alignment exhibit inconsistently, yielding a nearly random performance when faced contrasting instructions (Shen et al., 2023). Furthermore, LLMs that have been aligned through instruction tuning and RLHF remain susceptible to security compromises, even when further fine-tuned with only a minimal amount of harmful data (Qi et al., 2023). Also, merely fine-tuning with benign data does not ensure the model's security against such vulnerabilities (Qi et al., 2023).

Despite these limitations, it's essential to underline that while bottom-up approaches are successful in reaching basic alignment goals and mitigating attacks (Carlini et al., 2023), the formulation of these alignment goals remains intricate. The complexity is rooted in the moral values LLMs must embody, which are influenced by distinct contexts and user groups. Therefore, a nuanced consideration of these factors is essential during the deployment of LLMs.

4.2 Top-Down Approaches

Recent studies have shifted focus to top-down approaches. These approaches use explicit moral principles to align LLMs during inference and deployment, without modifying their parameters. This subsection systematically review these approaches.

4.2.1 Prompting & In-Context Learning

Ganguli et al. (2023) noted that LLMs could understand normative rules and follow instructions well, a capability that was enhanced with larger model sizes and training via RLHF. Following this, some studies augmented LLMs' morality by directly incorporating predefined standards into prompts (Zhang et al., 2023d; Xie et al., 2023). The application of Chain of Thought (CoT) prompting (Wei et al., 2022) bolster the moral reasoning of LLMs (Jin et al., 2022; Pan et al., 2023). Zhou et al. (2023b) proposed theory-guided prompting for moral judgement tasks. This approach is notable for embedding explicit moral theories within the prompts. Duan et al. (2023) employed LLMs to generate appropriate value instructions, which were then incorporated into prompts. Roy et al. (2023) leveraged LLMs to identify moral foundations and their associated roles (Roy et al., 2021) via few-shot in-context learning.

4.2.2 Output Refinement

Recent research have leveraged the self-evaluation and refinement capabilities of LLMs to enhance the morality of their output. Zhang et al. (2024) applied a two-step method that involved analyzing intentions and generating responses aligned with guidelines to prompt LLMs' self-improvement. Helbling et al. (2023) placed generated text into a custom prompt and used a separate LLM to assess if the text could be harmful. Ma et al. (2023) looked into using LLMs as ethical advisors, allowing them to refine their initial suggestions, especially if they were unethical. Li et al. (2023d) combined self-evaluation and rewind mechanisms, helping misaligned LLMs to independently produce answers in line with human ethical preferences through self-enhancement.

4.2.3 Activation Engineering

Li et al. (2023b) introduced inference-time intervention, a method that modifies model activations during inference by applying specific adjustments across a select number of attention heads, significantly enhancing LLaMA' performance (Touvron et al., 2023) on the TruthfulQA benchmark (Lin et al., 2021). Zou et al. (2023) developed RepE to extract representations linked to high-level concepts in LLMs such as honesty. Additionally, Li et al. (2023b) proposed InferAligner, employing safety steering vectors derived from safety-aligned LLMs to adjust the activations of base LLMs in response to harmful inputs.

4.2.4 Retrieval-Augmented Generation

Considering the diversity of morality across time and regions, Xu et al. (2023) employed Retrieval-augmented Generation (RAG) as a solution. Their

approach incorporates an external memory system, which archives key legal and moral principles, streamlining the alignment with evolving morality and enabling easy updates and customization. Pozzobon et al. (2023) applied RAG for toxicity-controlled text generation, accessing a dual datastore containing toxic and non-toxic samples to model text with respective undesirable and desirable attributes.

4.2.5 Others

Specifically, Takeshita et al. (2023) introduced the maximize expected choice (MEC) algorithm in response to the challenges of moral uncertainty.[3] This approach aggregates model outputs grounded in three normative ethical theories to generate the most appropriate response.

4.2.6 Discussion

The primary challenges LLMs encounter in moral tasks are (1) the diversity and metaphorical nature of morality; (2) the context-specific nature of LLM generation. Present top-down approaches aim to enhance LLMs mainly through prompting and context learning (Pan et al., 2023; Lin et al., 2023). However, achieving ethical alignment is not as straightforward as adhering to predefined ethical norms. We highlight that future studies should merge bottom-up and top-down methods, clearly articulating explicit moral values in application deployments.

5 Datasets and Benchmarks

This section reviews existing datasets and benchmarks utilized to evaluate morality in LLMs. Previous surveys have explored LLM evaluations concerning harmful content, bias, and misinformation (Chang et al., 2023; Sheng et al., 2021). Here, we concentrate specifically on the moral judgment and decision-making task.

5.1 Moral Norm Judgement

A variety of datasets have been developed to evaluate LMs' abilities in moral norm judgment. Notably, Emelin et al. (2020), Tay et al. (2020), Forbes et al. (2020), and Ziems et al. (2022) have each introduced the Moral Stories, MACS, SOCIAL CHEM-101, and MIC datasets, respectively. Moreover, Hendrycks et al. (2020) created the ETHICS dataset to underscore the challenges LMs face in ethical alignment with moral norms. Jin et al. (2022) launched the MoralExceptQA datasets, highlighting LLMs' struggles with flexible moral judgment. Jiang et al. (2021) trained Delphi model using CommonSense Norm Bank, a compilation that includes large-scale datasets like SOCIAL-CHEM101. Ziems et al. (2023) introduced NormBank dataset for assessing LLMs against situational social norms. Kiesel et al. (2022) offered a dataset spanning four geographical cultures aimed at classifying human values. Allein et al. (2023) introduced the interpretation modeling task accompanied by a dataset, designed to model multiple interpretations of a sentence's underlying semantics.

5.2 Moral Dilemma

Some current studies designed questionnaire tests for LLMs based on the moral machine experiment (Awad et al., 2018), which gathered moral decisions regarding autonomous vehicles in inevitable accident scenarios (Takemoto, 2024). Tanmay et al. (2023) and Khandelwal et al. (2024) utilized the Defining Issues Test dataset, which introduces a set of moral considerations on top of moral dilemmas, allowing LLMs to prioritize important factors before making decisions. Fränken et al. translated causal graphs into a prompt template, which is then used to procedurally generate a large and diverse set of moral dilemmas in a dataset called OffTheRails using a language model.

5.3 Generative Moral Decision-Making

Text-based games offer simulated environments rich in moral dilemmas. These game challenges LLMs with higher demands due to potential conflicts between game score and moral actions. While the Jiminy Cricket benchmark (Hendrycks et al., 2021) offers only binary moral assessments, the recent MACHIAVELLI benchmark (Pan et al., 2023) quantifies various harmful behaviors and evaluates agents across different moral dimensions. However, current LLM research in text-based games, relying on environment-provided action sets, does not fundamentally engage in generative tasks (Pan et al., 2023; Fang et al., 2024).

Beyond text-based games, Liu et al. (2023a) introduced the SANDBOX benchmark to evaluate LLMs' behavior through simulations of human society. Duan et al. (2023) developed DeNEVIL, an prompt generation algorithm that dynamically identifies and exploits the ethical vulnerabilities of

[3] *Moral Uncertainty* (Ecoffet and Lehman, 2021) refers to the state of being unsure about which moral theory is correct.

LLMs. Zhang et al. (2023b) introduced the VUM framework to evaluates LLMs' comprehension of more values by assessing the gap between discriminator critiques and human values. Ma et al. (2023) proposed ETHICSSUITE, a test suite that presents LLMs with complex and realistic moral dilemmas to assess their ethical reasoning capabilities.

6 Challenges and Future Directions

In this section, we discuss open problems and highlight important challenges for future work.

6.1 Challenges with Morality Evaluation

Existing works on assessing the moral capabilities of LLMs face certain limitations: (1) in moral norm judgement, the accuracy of advanced LLMs is approaching saturation. However, this does not accurately reflect the moral capabilities of LLMs. Future research should focus on the interpretability, consistency, and flexibility of LLMs in making moral judgments. (2) while existing studies demonstrate that LLMs exhibit variability across different cultures and languages (Khandelwal et al., 2024), systematic methods to measure these differences remain unexplored. (3) in generative moral decision-making, current benchmarks are under-researched. Text-based games offer a promising platform for such evaluations, but current studies rely heavily on the assumption of deriving oracle action sets. Future research could develop more open-world environments or eliminate the assumption of action dependency in current games.

6.2 Challenges with Moral Flexibility

A key challenge in LLM morality is the absence of a universally accepted moral standard, due to the changing nature of moral principles over time and their variation across cultures and languages. This issue is intensified by biases in the datasets used for training LLMs. Future efforts should prioritize the flexibility in moral reasoning. This involves (1) specifying clear ethical values in downstream tasks, and (2) ensuring that actions are appropriately aligned with the context and language.

6.3 Challenges with Moral Consistency

Another challenge lies in maintaining the consistency and robustness of LLMs' moral decision-making. Future research should include (1) designing metrics to measure moral consistency, taking into account that moral scenarios often lack clear-cut answers; (2) conducting evaluative work to explore the attributes of LLMs, such as the impact of fine-tuning extent and model size on moral inconsistencies.

6.4 Challenges with Moral Uncertainty

It is crucial to investigate how an LLM should act when uncertain about its course of action and what norms govern the uncertainty of normative statements. This issue is defined as *moral uncertainty*. Prior research in this area focused on RL-based agents (Ecoffet and Lehman, 2021), with recent efforts introducing the issue to LLMs (Takeshita et al., 2023). Given that LLMs possess knowledge of norms, exploring how they can leverage their reasoning capabilities in conjunction with tools or external feedback to make decisions under moral uncertainty is essential.

7 Conclusion

As LLMs increasingly impact our daily lives, it is crucial to explore their moral capacities to ensure they can perform automated tasks under appropriate human overseeing. This paper delves into the concept of morality within LLM research, highlighting specific downstream tasks related to morality, i.e., avoiding immoral generation, and moral judgement and decision-making. We review and summarize the performance of advanced LLMs on these tasks. We also review alignment techniques related to morality, including both bottom-up and top-down approaches. Furthermore, we review the datasets and benchmarks for assessing LLMs' moral capabilities. Lastly, we identify the primary challenges in this field and propose directions for future research.

Limitation

Considering the limitations, our review explores moral capabilities within LLMs. We focus on moral judgment and decision-making in downstream tasks. This emphasis is due to the existence of surveys that have already reviewed issues of bias, misinformation, and harmfulness in LLMs. Additionally, we concentrate on the current mainstream LLMs, discussing their key findings in moral tasks. For medium-scale LLMs like GPT-3, we provide a brief overview. Furthermore, we do not discuss the ethical performance of multimodal foundational models, as they present more complex issues not suitable for joint discussion with LLMs.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*.

David Abel, James MacGlashan, and Michael L Littman. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, volume 16, page 02. Phoenix, AZ.

Joshua Albrecht, Ellie Kitanidis, and Abraham J Fetterman. 2022. Despite" super-human" performance, current llms are unsuited for decisions about ethics and safety. *arXiv preprint arXiv:2212.06295*.

Liesbeth Allein, Maria Mihaela Truşcă, and Marie-Francine Moens. 2023. Interpretation modeling: Social grounding of sentences by reasoning over their implicit moral judgments. *arXiv preprint arXiv:2312.03726*.

Guilherme FCF Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2023. Exploring the psychology of gpt-4's moral and legal reasoning. *arXiv preprint arXiv:2308.01264*.

Michael Anderson and Susan Leigh Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4):15–15.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2022. Enabling classifiers to make judgements explicitly aligned with human values. *arXiv preprint arXiv:2210.07652*.

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52.

Paul Bello and Bertram F Malle. Computational approaches to morality.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Vamshi Krishna Bonagiri, Sreeram Vennam, Manas Gaur, and Ponnurangam Kumaraguru. 2024. Measuring moral inconsistencies in large language models. *arXiv preprint arXiv:2402.01719*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. 2023. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Canyu Chen and Kai Shu. 2023a. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Canyu Chen and Kai Shu. 2023b. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Paul Conway and Bertram Gawronski. 2013. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, 104(2):216.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023a. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2023b. Probing explicit and implicit gender bias through llm conditional text generation. *arXiv preprint arXiv:2311.00306*.

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Tun Lu, Xing Xie, and Ning Gu. 2023. Denevil: Towards deciphering and navigating the ethical values of large language models via instruction learning. *arXiv preprint arXiv:2310.11053*.

Adrien Ecoffet and Joel Lehman. 2021. Reinforcement learning under moral uncertainty. In *International conference on machine learning*, pages 2926–2936. PMLR.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. *arXiv preprint arXiv:2401.09334*.

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087*.

Erez Firt. 2023. Ought we align the values of artificial moral agents? *AI and Ethics*, pages 1–10.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Jan-Philipp Fränken, Ayesha Khawaja, Kanishk Gandhi, Jared Moore, Noah D Goodman, and Tobias Gerstenberg. Off the rails: Procedural dilemma generation for moral reasoning.

Kathleen C Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. *arXiv preprint arXiv:2205.12771*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Bernard Gert and Joshua Gert. 2002. The definition of morality.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 281–291.

Thilo Hagendorff and David Danks. 2023. Ethical and methodological challenges in building morally informed ai systems. *AI and Ethics*, 3(2):553–566.

Ahmed Abdeen Hamed and Xindong Wu. 2023. Improving detection of chatgpt-generated fake science using real publication text: Introducing xfakebibs a supervised-learning network algorithm.

Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. 2023. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021. What would jiminy cricket do? towards agents that behave morally. *arXiv preprint arXiv:2110.13136*.

Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023a. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023b. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023c. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

Shelly Kagan. 2018. *Normative ethics*. Routledge.

Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. *arXiv preprint arXiv:2402.02135*.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023a. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023b. The moral authority of chatgpt. *arXiv preprint arXiv:2301.07098*.

Joe Kwon, Sydney Levine, and Joshua B Tenenbaum. 2023. Neuro-symbolic models of human moral judgment: Llms as automatic feature extractors. In *Workshop on Challenges of Deploying Generative AI: 40th International Conference on Machine Learning*.

Nathan Lambert and Roberto Calandra. 2023. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023c. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2023d. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Kush R Varshney. 2023. Towards healthy ai: Large language models need therapists too. *arXiv preprint arXiv:2304.00416*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023a. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Pingchuan Ma, Zongjie Li, Ao Sun, and Shuai Wang. 2023. " oops, did i just say that?" testing and repairing unethical suggestions of large language models

with suggest-critique-reflect process. *arXiv preprint arXiv:2305.02626*.

David Mhlanga. 2023. Open ai in education, the responsible and ethical use of chatgpt towards lifelong learning. *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)*.

Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsu H Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *arXiv preprint arXiv:2310.19677*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR.

Mark Pock, Andre Ye, and Jared Moore. 2023. Llms grasp morality in concept. *arXiv preprint arXiv:2311.02294*.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5108–5125.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2023. Towards few-shot identification of morality frames using in-context learning. *arXiv preprint arXiv:2302.02029*.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. *arXiv preprint arXiv:2109.04535*.

Chandan Kumar Sah, Dr Lian Xiaoli, and Muhammad Mirajul Islam. 2024. Unveiling bias in fairness evaluations of large language models: A critical literature review of music and movie recommendation systems. *arXiv preprint arXiv:2401.04057*.

Ari Saptawijaya and Luís Moniz Pereira. 2016. Logic programming for modeling morality. *Logic Journal of the IGPL*, 24(4):510–525.

Nino Scherrer, Claudia Shi, Amir Feder, and David M Blei. 2023. Evaluating the moral beliefs encoded in llms. *arXiv preprint arXiv:2307.14324*.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. 2019. Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*.

Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in agi safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. 2023. The trickle-down impact of reward (in-) consistency on rlhf. *arXiv preprint arXiv:2309.16155*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Zijing Shi, Meng Fang, Yunqiu Xu, Ling Chen, and Yali Du. 2022. Stay moral and explore: Learn to behave morally in text-based games. In *The Eleventh International Conference on Learning Representations*.

Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

Walter Sinnott-Armstrong. 2003. Consequentialism.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023a. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023b. Salmon: Self-alignment with principle-following reward models. *arXiv preprint arXiv:2310.05910*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023c. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Kazuhiro Takemoto. 2024. The moral machine experiment on large language models. *Royal Society Open Science*, 11(2):231393.

Masashi Takeshita, Rzepka Rafal, and Kenji Araki. 2023. Towards theory-based moral ai: Moral ai with aggregating models based on normative ethical theory. *arXiv preprint arXiv:2306.11432*.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *arXiv e-prints*, pages arXiv–2309.

Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Christopher Pal. 2020. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5369–5373.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2023. How prevalent is gender bias in chatgpt?–exploring german and english chatgpt responses. *arXiv preprint arXiv:2310.03031*.

Aleksandra Urman and Mykola Makhortykh. 2023. The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554.

Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wendell Wallach, Colin Allen, and Iva Smit. 2020. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In *Machine Ethics and Robot Ethics*, pages 249–266. Routledge.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Changyu Wang, Siru Liu, Hao Yang, Jiulin Guo, Yuxuan Wu, and Jialin Liu. 2023a. Ethical considerations of using chatgpt in health care. *Journal of Medical Internet Research*, 25:e48009.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Kimi Wenzel, Geoff Kaufman, and Laura Dabbish. 2023. Beyond fairness: Alternative moral dimensions for assessing algorithms and designing systems. *arXiv preprint arXiv:2312.12559*.

Jiaying Wu and Bryan Hooi. 2023. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. *arXiv preprint arXiv:2310.10830*.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, pages 1–11.

Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. 2023. Align on the fly: Adapting chatbot behavior to established norms. *arXiv preprint arXiv:2312.15907*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. 2023a. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values. *arXiv preprint arXiv:2311.10766*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023b. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023c. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jentse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023a. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023b. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Yuqi Zhang, Liang Ding, Lefei Zhang, and Dacheng Tao. 2024. Intention analysis prompting makes large language models a good jailbreak defender. *arXiv e-prints*, pages arXiv–2401.

Zhaowei Zhang, Fengshuo Bai, Jun Gao, and Yaodong Yang. 2023b. Measuring value understanding in language models through discriminator-critique gap. *arXiv preprint arXiv:2310.00378*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023c. Safety-bench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Zhexin Zhang, Junxiao Yang, Pei Ke, and Minlie Huang. 2023d. Defending large language models against jailbreaking attacks through goal prioritization. *arXiv preprint arXiv:2311.09096*.

Jiachen Zhao, Zhun Deng, David Madras, James Zou, and Mengye Ren. 2023a. Learning and forgetting unsafe examples in large language models. *arXiv preprint arXiv:2312.12736*.

Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. 2023b. Gptbias: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023c. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023b. Rethinking machine ethics–can llms perform moral reasoning through the lens of moral theories? *arXiv preprint arXiv:2308.15399*.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023c. Making harmful behaviors unlearnable for large language models. *arXiv preprint arXiv:2311.02105*.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.