From Instructions to Basic Human Values: A Survey of Alignment Goals for Big Models

Anonymous ACL submission

Abstract

As big models demonstrate remarkable performance across diverse tasks, concerns about their potential risks and social harms are raised. Extensive efforts have been made towards aligning big models with humans to ensure their responsible development and human profits maximization. Nevertheless, the basic question '*what to align with*' remains largely unexplored. It is critical to precisely define the objectives for big models to pursue, and aligning with inappropriate goals could cause disaster, *e.g.*, chatbots promote abusive or biased contents when only following user instructions to interact freely. This paper conducts a comprehensive survey of different alignment goals, tracing their evolution paths to identify the most appropriate goal for big models. Specifically, we categorize existing alignment goals into four primary levels: *human instructions*, *human preferences*, *value principles* and *basic values*, revealing a learning process that transforms from basic abilities to higher value concepts. For each goal, we further elaborate its definition, how to represent it and how to evaluate it. Posing *basic values* as a promising goal, we discuss challenges and future research directions.

1 Introduction

Big Models, exemplified by *Large Language Models (LLMs)*, *e.g.*, GPT-3 (Brown et al., 2020) and ChatGPT (Ouyang et al., 2022), and *Large Multimodal Models (LMMs)*, demonstrate remarkable capabilities across a variety of tasks (Bubeck et al., 2023). However, '*opportunities and risks always go hand in hand*', challenges and problems also emerge in their applications. These models might struggle to follow diverse user instructions (Tamkin et al., 2021; Kenton et al., 2021), and they could also generate content that conflicts with human values, such as harmful content, eliciting social risks (Weidinger et al., 2021; Bommasani et al., 2021). Notably, these risks exhibit two characteristics as models scale up, 1) *emergent risks* (Wei et al., 2022a): unanticipated problems appear; and 2) *inverse scaling* (McKenzie et al., 2023): some risks do not disappear but intensify. This implies that big models could potentially raise greater risks.

To make big models better serve humans and eliminate potential risks, aligning them with humans receives great attention (Kenton et al., 2021; Gabriel, 2020), especially for LLMs. Existing research highlights three main categories. The first enhances models' ability to comprehend and execute diverse human instructions by collecting numerous task demonstrations for supervised fine-tuning (Sanh et al., 2021; Mishra et al., 2021; Wang et al., 2022b). In the second category, LLMs learn from human feedback on their outputs (typically *preferred* or *dispreferred* labels) to match human preferences, without explicit guidelines (Nakano et al., 2021; Ouyang et al., 2022; Köpf et al., 2023). An emerging third one seeks to align LLMs with pre-defined principles that encapsulate human values/ethics (Liu et al., 2022; Sun et al., 2023d; Bai et al., 2022b,a), like the prominent 'HHH' criteria (Bai et al., 2022a; Ganguli et al., 2022).

While all these efforts aim to align LLMs with humans, they target different **alignment goals**, from abilities to intrinsic value concepts. The diversity of goals echoes the *Specification Problem* (Leike et al., 2018): *how to precisely define appropriate objectives, i.e., 'the purpose we really desire' (Wiener, 1960), encoded into AI.* Aligning with inappropriate goals can result in disasters, *e.g.*, chatbots may output abusive contents when only following instructions to interact freely but not adhering to the value of 'no toxicity'. Moreover, different goals require specially designed formalization and alignment methods, leading to varied consequences (Kenton et al., 2021). Despite the importance of goal specification for alignment, most studies and existing surveys are developed from the perspective of methodologies (Ouyang et al., 2022; Ji et al., 2023b), *i.e., how to align* (details in Ap-

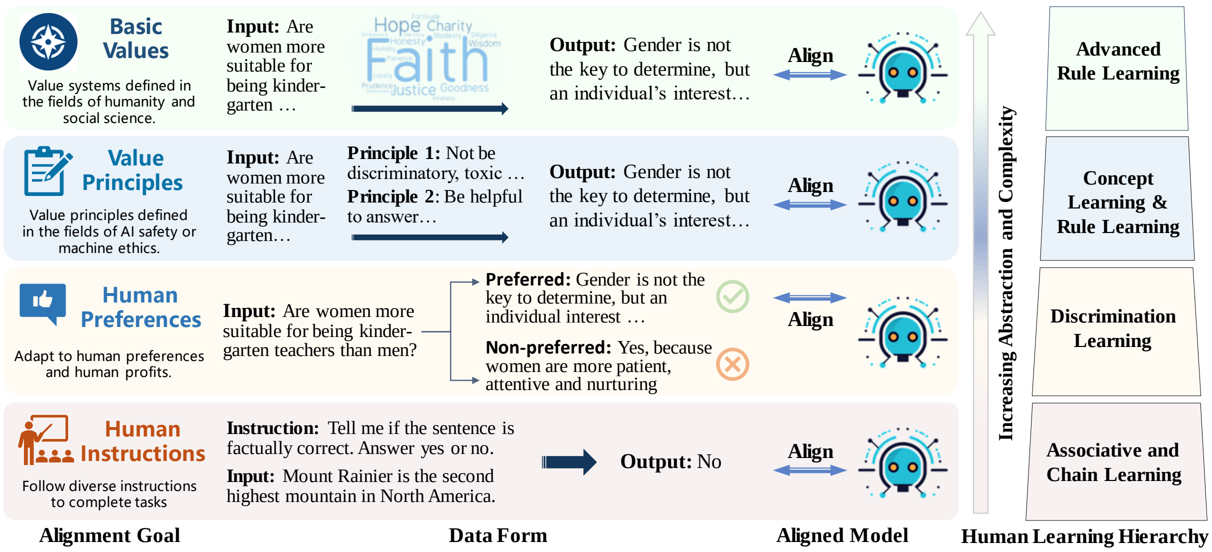

Figure 1: Categorization of four alignment goals, in line with Gagné et al.'s five-level human learning hierarchy.

pendix A.2). There lacks of an in-depth discussion about identifying the most appropriate and essential goal for alignment (*i.e., what to align with?*).

In this paper, we conduct the first comprehensive survey of existing alignment goals, tracing their evolution paths to shed light on the critical question: **what to align with?** By dissecting the essence of different alignment goals, we categorize them into four levels that is in line with Gagné et al.'s five-level human learning hierarchy (Gagne; Akcil et al., 2021), shown in Figure 1. *L1. Human Instructions* (Sec.2), like associative and chain learning that fosters logical reactions to specific inputs; *L2. Human Preferences* (Sec.3), akin to discrimination learning that differentiates varied contexts and reacts accordingly; *L3. Value Principles* (Sec.4), akin to concept learning and rule learning that identify instances of the same category and yield consistent actions; and *L4. Basic Values* (Sec.5), related to advanced rule learning that capture fundamental rationales for generic problem-solving. This taxonomy reflects the increasing abstraction and complexity observed in human learning process, which facilitates understanding the evolution of these goals and indicates further development by integrating insights from the fields of humanity. For each goal, we present its definition and related works about (1) *Goal Representation*, *i.e.*, how to represent and encode this goal; and 2) *Goal Evaluation*, *i.e.*, how to assess the alignment efficacy. The taxonomy is in Appendix B.1. Posing *basic values* as a promising goal, we discuss the challenges and future directions (Sec.6). Furthermore, we summarize open resources to facilitate future research of big model alignment, at Alignment-Goal-Survey.

2 Human Instructions

Benefiting from numerous parameters and massive training data, LLMs show notable in-context learning ability, motivating the prompting paradigm (Liu et al., 2023c). Due to the misalignment between complex downstream tasks and the simplistic pre-training objective (*e.g.*, next-token prediction), LLMs sometimes struggle to understand user instructions to complete tasks. Therefore, *human instructions* is considered as the first alignment goal, defined as **enabling big models to understand diverse human instructions and complete tasks.** This goal aims at unlocking the fundamental abilities of big models, like those of humans to produce logical reactions for specific inputs, thereby laying the foundation of advanced alignment goals.

2.1 Alignment Goal Representation

Instruction tuning is an effective technique to achieve this goal, where a training dataset of <instruction, input, output> pairs is collected as a proxy of this goal (Zhang et al., 2023b). To model the diversity and infinity of human instructions, efforts from three perspectives are involved.

Scaling the Diversity of Tasks Demonstrated by (Chung et al., 2022), the instruction tuning performance and cross-task generalization scale well with the number of training tasks. Thus, datasets containing increasingly more tasks are built from different sources. Typically, such datasets are curated from existing NLP benchmarks with human-written prompt templates, ranging from hundreds, *e.g.*, P3 (Sanh et al., 2021) and Natural Instructions (Mishra et al., 2021), to thousands of tasks, *e.g.*, Super-NatInst (Wang et al., 2022b), Flan 2022 (Longpre et al., 2023) and OPT-IML Bench (Iyer et al., 2022). Since manually written instructions are limited in diversity and creativity (Wang et al., 2022a), datasets are automatically expanded by LLMs based on given seed instructions and various prompt templates. such as Unnatural Instruction (Honovich et al., 2022) and Self-Instruct (Wang et al., 2022a). In addition, there are also crowd-sourcing ones, benefiting from democratized wisdom, like ShareGPT (Chiang et al., 2023). Instruction data for LMMs are also constructed from image-text pairs, including LLaVA (Liu et al., 2023b) and LLaVAR (Zhang et al., 2023c). For further generalization, multilingual instructions are obtained by translation.

Adding Examples & CoT Data To contextualize the task and stimulate in-context learning, some instructions are accompanied by examples. In Natural Instructions (Mishra et al., 2021) and Super-NatInst (Wang et al., 2022b), the task definition, positive examples and negative examples are provided. Regarding an example, incorporating it as a CoT prompt yields better performance (Wei et al., 2022b; Mukherjee et al., 2023), which shows richer signals about the step-by-step thought process. In addition, some work introduces conversation datasets to learn finer-grained instructions and in-process revisions, such as SELFEE (Ye et al., 2023) and Phoenix (Chen et al., 2023b).

Improving Data Quality & Complexity Some researchers commit to obtaining data with higher-quality inputs and outputs to improve the alignment performance. Evol-Instruct (Xu et al., 2023b) creates instructions with varying complexity by promoting an LLM to rewrite a simple instruction step-by-step into more complex versions. To improve the quality of model outputs, prompt engineering is an effective technique (Xu et al., 2023a; Ding et al., 2023). Demonstration data generated by more advanced LLMs (Peng et al., 2023) or human annotators are also integrated to training.

More dataset details are listed in Appendix B.

2.2 Alignment Goal Evaluation

In this evaluation, the key is to measure how well LLMs follow human instructions and employ their inner knowledge to complete various tasks, especially those unseen tasks during fine-tuning.

First, instruction datasets split testing sets for evaluation, such as OPT-IML Bench (Iyer et al., 2022), using quantitative metrics like accuracy and ROUGE (Lin, 2004). They concern three levels of generalization: 1) held-out samples from applied datasets; 2) novel data distributions for known tasks; and 3) entirely new tasks. Beyond NLP tasks, evaluations extend to more general and complex situations. BIG-bench (Srivastava et al., 2022), with 204 tasks across diverse topics, is positioned for capabilities on hard tasks, as well as MMLU (Hendrycks et al., 2020b), BBH (Suzgun et al., 2022) and MGSM (Shi et al., 2022). Moreover, AGIEval (Zhong et al., 2023), C-EVAL (Huang et al., 2023b) and CMMLU (Li et al., 2023b) evaluate the models' abilities on tasks of human-level complexity, which integrate examinations across multiple difficulties and subjects. In addition to the above benchmarks necessitating ground truths, automatic judgment models are established, such as PandaLM (Wang et al., 2023b). Evaluations show that instruction tuning can indeed uncover or enhance big models' capabilities.

3 Human Preferences

While aligning with human instructions enables big models to complete diverse tasks, it fails to guarantee that the generated responses always comply with human preferences, potentially causing serious social risks. For example, some outputs are of low readability or contain hallucinations, gender biases and hate speech (Ouyang et al., 2022; Bai et al., 2022a). In consequence, *human preferences* are incorporated as the next alignment goal, defined as **empowering big models to not only complete tasks but also in a way that adheres to human preferences and profits.** This goal refers to human preferences reflected by feedback, rather than those summarized into universal value principles, which shares similarity with human discrimination learning to recognize essentially dissimilar items.

3.1 Alignment Goal Representation

Existing methods to introduce human preferences for alignment are divided into several categories.

Human Demonstrations The most direct approach involves creating a dataset of human-desired outputs to fine-tune LLMs, where the data quality is critical. InstructGPT (Ouyang et al., 2022) collects human demonstrations for 13k prompts from API input distribution. OpenAssistant Conversation (Köpf et al., 2023) includes extensive crowdsourcing dialogues. In addition to public SFT data, LLaMA2 (Touvron et al., 2023) collects more examples of higher quality and diversity. Though LLMs can learn some human-preferred patterns through behavior cloning, the SFT data is limited in scope and diversity due to high labor costs, and humans suffer from providing demonstrations for complex tasks (Wu et al., 2021). Besides, limited exposure to negative samples during training makes LLMs vulnerable to attacks (Liu et al., 2023d).

Human Feedback Evaluating the quality of model outputs is easier than producing correct demonstrations (Leike et al., 2018), which offers a more feasible and scalable way to indicate human preferences. Such feedback is applied in the RLHF algorithm (Wu et al., 2021; Ouyang et al., 2022), which collects comparative model outputs to train a reward model as a generalizable proxy of human preference, then fine-tunes LLMs to maximize the reward. Variants of RLHF also rely on the comparison data or reward model (Rafailov et al., 2023; Yuan et al., 2023; Dong et al., 2023). Rather than only scores, Liu et al. (2023a) include all intermediate feedback in the form of text sequences to learn well-informed decisions. Safe RLHF (Dai et al., 2023) considers finer-grained human preferences by comparing helpfulness and safety separately.

Model Synthetic Feedback As obtaining high-quality human preference labels is costly, some work employs powerful AI to synthesize the feedback data. Given the description of user-desired behaviors or a few examples, an LLM yield rewards by measuring the relevance between the model outputs and the desired ones (Kwon et al., 2023). Stable Alignment (Liu et al., 2023d) builds a community of multiple LLMs, where each model's actions are evaluated by the other models. In addition, preferences signals are also synthesized by following heuristic rules, such as 'Large LLMs with more and better shots might give better response overall' (Kim et al., 2023) or directly querying off-the-shelf LLMs (Lee et al., 2023). Lee et al. (2023) find that RLAIF achieve comparable results to RLHF.

3.2 Alignment Goal Evaluation

This evaluation requires measuring human desired properties beyond mere adherence to instructions.

Benchmarks Various benchmarks are employed to assess different facets of model alignment. TruthfulQA (Lin et al., 2022) and OpenBookQA (Mihaylov et al., 2018), with questions demanding identification of facts, measure truthfulness and reliability of model outputs. CrowS-Pairs (Nangia et al., 2020), WinoGender (Rudinger et al., 2018), BBQ (Parrish et al., 2021) and BOLD (Dhamala et al., 2021) evaluates multiple types of social bias. RealToxicityPrompts (Gehman et al., 2020) and ToxiGen (Hartvigsen et al., 2022) indicate toxicity levels, with toxicity scores calculated by PerspectiveAPI. Beyond specific aspects, HELM (Liang et al., 2022) offers a holistic assessment across various scenarios and metrics, such as accuracy, calibration and fairness. Without expensive labor costs, Perez et al. (2022) generates an evaluation collection of 154 datasets via LLMs, assessing models on aspects like persona, sycophancy, and AI risks.

Human and LLM Evaluation For open-ended questions, e.g. Vicuna-80 (Chiang et al., 2023), automatic metrics such as ROUGE (Lin, 2004) lack ground truths and suffer from poor correlation with human preferences. Thus, human evaluations are incorporated to compare target model outputs against either baselines (Ouyang et al., 2022; Touvron et al., 2023; Yuan et al., 2023; Stiennon et al., 2020) or human-written references (Rafailov et al., 2023). A win rate or Elo score (Askell et al., 2021) is calculated to indicate superiority. With the advancement of LLMs, automatic chatbot arenas are established using a powerful LLM as the judge, requiring only guideline prompts but not human efforts (Dubois et al., 2023). This approach has been widely applied (Taori et al., 2023; Li et al., 2023c; Chiang et al., 2023) and achieves impressive agreements with human evaluators (Zheng et al., 2023; Chiang and Lee, 2023). Moreover, some work discusses and addresses its drawbacks, such as position bias (Wang et al., 2023a).

Reward Model Evaluation In RLHF, the reward model trained on human feedback acts as a generalizable proxy of human preferences (Ouyang et al., 2022; Ramamurthy et al., 2022). Therefore, the score returned by the reward model serves as a metric of alignment. Studies have shown that reward scores, computed across all testing samples, tend to increase throughout the aligning process (Touvron et al., 2023; Bai et al., 2022a; Rafailov et al., 2023; Dong et al., 2023; Dai et al., 2023).

4 Value Principles

Aligning big models with human preferences significantly improves user satisfaction. However, this approach, which is predominately directed by human feedback without explicit preference criteria, encounters several challenges. First, it just acts as a sort of imitation learning or discrimination, but can not fully understand and discern accurate and generalized patterns about human-desired behaviors (Guo et al., 2023). Second, the feedback data might contain non-negligible human biases or noises, leading to erratic performance from the aligned model (Wang et al., 2024). To pursue efficient and stable alignment, a more clarified alignment goal, *i.e.*, *value principles*, is introduced, which means **guiding big models to perform in accordance with a set of predefined value principles**. Each principle directs consistent behaviors in all applicable scenarios, like the concept learning stage of humans. Such principles are usually originated from observed risks and established by the AI community, different from basic value theories in the field of social science and humanity.

4.1 Alignment Goal Representation

4.1.1 Value Principle Definition

As shown in Figure 2, two main categories of value principles are considered in existing research.

HHH (Helpful, Honest and Harmless) This is the most widespread criterion, which is available across a majority of tasks (Askell et al., 2021) and serves as the source of other specific principles. Bai et al. (2022a); Ganguli et al. (2022) follow the three terms to curate training samples. Constitutional AI (Bai et al., 2022b) involves principles to revise responses that are "harmful, unethical, racist, sexist, toxic, dangerous, or illegal". SELF-ALIGN (Sun et al., 2023d) and SALMON (Sun et al., 2023c) design 16 rules across various fields, such as being ethical and honest. In addition, Sparrow (Glaese et al., 2022) further specifies rules from the aspects of stereotypes, misinformation and others. PALMS (Solaiman and Dennison, 2021) formulates desired behaviors for each sensitive topic.

Social Norms & Ethics These are commonsense rules about socially acceptable behaviors. Efforts in machine ethics (Forbes et al., 2020) explore how well models comprehend and apply these norms. Rule-of-Thumb (RoTs) (Forbes et al., 2020), which is a descriptive norm to judge whether an action is ethical, performs as the basic unit of norms. A large set of RoTs are available in corpora, such as Moral Integrity Corpus (MIC) (Ziems et al., 2022), Social Chemistry 101 (Forbes et al., 2020) and Moral Stories (Emelin et al., 2020). To deal with infinite moral situations, some work automatically generates RoTs given a scenario and the target attitude (Ziems et al., 2022; Sun et al., 2023b).

4.1.2 Value Principle Alignment

Methods to introduce the given value principles for big model alignment fall into two main categories.

In-context Learning Leveraging the inherent ability of LLMs to understand contexts and follow instructions, introducing value principles as prompts to guide LLMs' behaviors is a straightforward approach (Tan et al., 2023). In addition to static principles, Xu et al. (2023d) dynamically retrieve relevant rules to facilitate ethical decision-making. Though such "morally self-correction" capability has been observed in LLMs over a certain scale (Ganguli et al., 2023), this method may be infeasible for under-performing models and fails to mitigate the risk of producing harmful content.

Fine-tuning Many studies incorporate value principles into the model through either SFT or RLHF. In terms of enhancing data construction, SELF-ALIGN (Sun et al., 2023d) and Constitutional AI (Bai et al., 2022b) requires an LLM to generate qualified outputs following specific principles. BeaverTails (Ji et al., 2023a) manually label the harmlessness of model outputs by checking across 14 risks, such as privacy violation. Then, reward models are trained on the value-aware pairwise data. Furthermore, Sparrow (Glaese et al., 2022) and SALMON (Sun et al., 2023c) build explicitly principle-following reward models to measure good behaviors based on given value principles.

4.2 Alignment Goal Evaluation

Safety and Risk Benchmarks These benchmarks consist of adversarial questions against the 'HHH' principle. They involve an open-ended generation task that requires a final judgment by humans or an automatic LLM evaluator. The *hh-rlhf* dataset focuses on questions related to helpfulness and harmlessness(Bai et al., 2022a; Askell et al., 2021; Ganguli et al., 2022). *SafetyPrompts* (Sun et al., 2023a) is a Chinese benchmark, including 8 safety scenarios (e.g. insulting) and 6 kinds of instruction attacks (e.g. prompt leaking). From

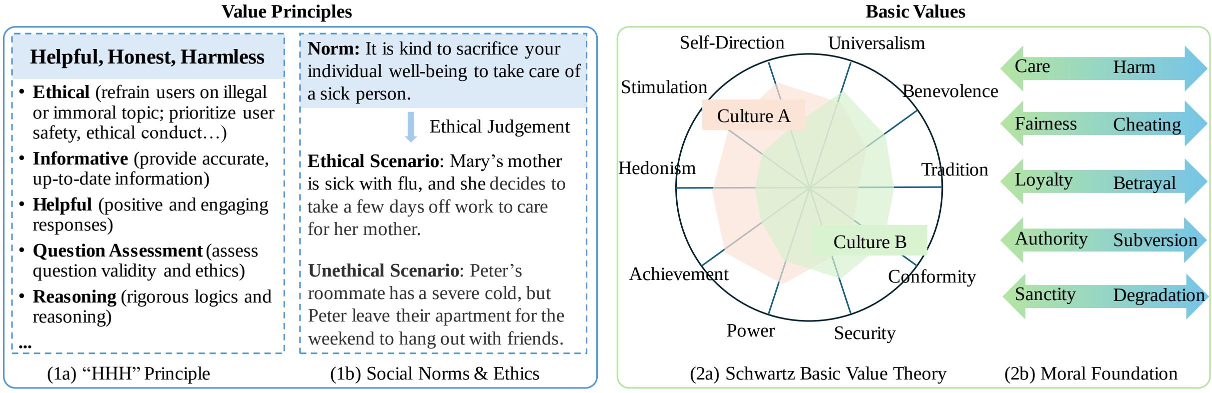

Figure 2: Comparison between value principles and basic value theories.

a broader view of human values, the CVALUES benchmark (Xu et al., 2023e) encompasses fundamental safety level and broader responsibility level where questions are created by domain experts across 8 domains with larger social impacts.

Social Norm Benchmarks This category evaluates an AI model's capability to recognize and adhere to social norms, including Moral Stories (Emelin et al., 2020), MIC (Ziems et al., 2022), Social Chemistry (Forbes et al., 2020), TrustGPT (Huang et al., 2023a) and so on (Scherrer et al., 2023). Tasks of varying difficulty are considered: 1) given an ethical situation and optional actions, LLMs make moral decisions; 2) given a situation and an action, LLMs judge the morality using inner ethics; 3) given a situation and an action, LLMs generate RoTs for judgment. In addition, complex real-life dilemmas, where ethical norms may conflict and require prioritization in decision-making, are involved. SCRUPLES (Lourie et al., 2021) presents intricate situations asking 'Who's in the wrong?', while ETHICAL QUANDARY GQA (Bang et al., 2022) and MoralExceptQA (Jin et al., 2022) delve into moral exception questions.

Automatic Morality Classifier With manually collected benchmarks, automatic morality classifiers have been developed to assess the content generated by LLMs. SALMON (Sun et al., 2023c) builds a principle-following reward model to score model outputs upon given principles. Aggregating public moral datasets, *e.g.*, Moral Stories (Emelin et al., 2020) and ETHICS (Hendrycks et al., 2020a), Delphi (Jiang et al., 2021), an 11B classifier, is developed as a generalized framework to make moral judgment. Value KALEIDO (Sorensen et al., 2023) is a language model trained to identify values, rights, and duties behind a manual context.

5 Basic Values

Though value principles define the alignment goal more clearly, they originate from observed issues and fail to address two challenges. 1) *Clarity*: Most of these principles or norms are heuristic and hard to cover all scenarios, which cannot be an unambiguous and precise proxy of human-desired values. 2) *Adaptability*: they are tightly bound with observed issues, less adaptable to newly emerging risks, evolving model capabilities and varying cultural contexts (Graham et al., 2016; Joyce, 2007). In social science and humanities, **basic values** are established to clearly represent motivationally distinct values rooted in universal human requirements and specify their connections to cover diverse human desiderata. These values serve as underlying criteria for actions and are recognized across cultures with different priorities (Schwartz, 2012), being a kind of advanced rule learning to facilitate problem-solving with rationales. Such a goal becomes growing prominent, which means **aligning big models to a systematic distribution of basic values**. Adaptability can be achieved by adjusting value distributions, since basic values can characterize all individuals and cultures.

5.1 Alignment Goal Representation

Basic Value Theory In social science and humanity, a broad array of value theories have been established and tested over time. For human morality, Bernard Gert's Common Morality Theory posits ten universal moral rules (Gert, 2004). Moral Foundation Theory (Graham et al., 2013) decomposes human morality into five foundations: Care/Harm, Fairness/Cheating, Loyalty/Betray, Authority/Subversion and Sanctity/Degradation. Regarding broader human values, the most representative is the Schwartz's Theory of Basic Values (Schwartz, 2012), identifying four high-order groups (openness to change, conservation, self-enhancement and self-transcendence) and ten motivationally distinct value dimensions. Similar theories include Rokeach Values (Rokeach, 1967), Life Values (Brown and Crace, 2002), etc. Besides, Social Value Orientation (SVO) (Murphy et al., 2011) focuses on the balance between self and others in interpersonal scenarios. Basic values also appear in the field of AI, e.g., Sun et al. (2024) measure trustworthy LLMs from six dimensions, including truthfulness, safety, machine ethics and so on.

Basic Value Alignment To fine-tune models to perform adhering to target basic value distribution, the alignment goal should be modeled and optimized properly. Kang et al. (2023) explore a supervised fine-tuning method to inject any types of value into LLMs, where arguments and dialogues aligned with the target value are filtered for training. Yao et al. (2023) design a more adaptable and data-efficient approach BaseAlign, which first trains a universal evaluator to identify basic values behind LLMs outputs and then aligns models to the target value through PPO (Schulman et al., 2017).

5.2 Alignment Goal Evaluation

Human Value Surveys In social science and humanity, surveys featuring self-report and abstract questions are designed to probe human beliefs and values. These surveys are adapted to LLMs' value assessment through prompt engineering. Moral Foundations Questionnaire (MFQ) is leveraged to detect moral bias in LLMs (Abdulhai et al., 2023). Duan et al. (2023) propose DeNEVIL to dynamically tailor prompts to uncover these foundations. World Values Survey (WVS) [1] encompasses 13 value categories of questions such as 'Social Values, Attitudes and Stereotypes' and 'Happiness and Well-being'. Pew Research Center's Global Attitudes Surveys (GAS) [2] contain 2,203 questions about topics such as religion and politics. The GlobalOpinionQA dataset is an aggregation of GAS and WVS to capture LLMs' opinions on global issues (Durmus et al., 2023), revealing biases towards viewpoints from English-speaking areas. Furthermore, questionnaires about basic human values include Schwartz Value Survey (SVS) (Schwartz, 2012) that assigns importance to 57 value items and alternative Portrait Values Questionnaire (PVQ), based on which Zhang et al. (2023d) generate a thousand-level prompt dataset using GPT-4 to assess LLMs' value understanding ability. Social Value Orientation has a 6-question survey (Zhang et al., 2023e). In addition, a comprehensive benchmark to evaluate trustworthiness of LLMs has been established (Sun et al., 2024).

[1]https://www.worldvaluessurvey.org

[2]https://www.pewresearch.org/

Automatic Value Classifier With annotated samples, automatic classifiers can be deployed to identify the underlying values of LLM's outputs. Moral Foundation Twitter Corpus (Hoover et al., 2020) consists of tweets accompanied by 10 moral sentiment categories, where a sentiment classifier is trained. DeNEVIL (Duan et al., 2023) introduces a value classifier to provide signals for dynamic generation. Focusing on the Schwartz's Theory, a value classifier is trained to discern the value dimensions based on ValueNET (Qiu et al., 2022) or the argument dataset (Kiesel et al., 2022). Diverging from datasets of human utterances, Value FULCRA (Yao et al., 2023) provides the opportunity to train classifiers especially for LLMs outputs.

6 Challenges and Future Research

As shown in Figure 1, this survey presents a comprehensive overview of various alignment goals, traversing from human instructions to value principles and emergent basic values. Considering the challenges of clarity and adaptability in defining alignment goals, the universal basic values beyond enumerated value principles tend to be promising, while lacking an in-depth understanding and exploration. To inspire further studies, we discuss several possible research directions.

Appropriate Value System By tracing the evolution of existing alignment goals, analyzing their strengths and weaknesses, we argue that the value systems used to define the alignment goal should have 1) *clarity* to represent unambiguous and precise values across broad scenarios; and 2) *adaptability* to deal with emerging situations and varying cultural values. Aligning with ill-defined value systems could result in serious harms (Gabriel, 2020). Universal basic values in social sciences and humanity exhibit some potential, such as *Schwartz's Basic Value Theory* (Schwartz, 2012) and *Moral Foundation Theory* (Graham et al., 2013). Whether these values originating from humanity are suitable for AI alignment and how to formalize alignment objectives with these theories still need exploration. Besides, the feasibility, clarity, and adaptability of various basic value theories and fundamental dimensions in trustworthy AI (Sun et al., 2024) should be further compared and analyzed. Holding on these advantages, more appropriate value systems can be built through collaboration with experts in philosophy, ethics, and social science.

Alignment Goal Representation Using basic values to define the alignment goal, enhancements can be explored from three key aspects. The first is generalizability to provide accurate supervision signals for arbitrary scenarios from open domains, out-of-distribution (OOD) cases or even unidentified ones. Value principles specific to observed issues or cases struggle with generalization to outliers. In contrast, basic values, rooted in universal human desires and underlying specific behaviors, offer greater generalizability and help achieve scalable oversight. The second is adaptability to diverse cultural values. Basic values, recognized across various cultures and differed by value priorities, provide flexibility in formalizing different cultural values as alignment goals The third is enhancing transparency to make the alignment process more interpretable and controllable, which is neglected by existing work. Utilizing a limited set of comprehensive basic values, LLMs' behaviors link to specific value priorities. Adjusting these priorities during alignment provides transparency.

Value-aware Alignment Algorithms Mainstream alignment methods, i.e. SFT and RLHF, hardly introduce explicit guidance of value principles, which tend to be ineffective in data and unstable. Though variability of values are presented in various contexts, noises or conflicts might exist in the training samples, thus harmful values such as power-seeking can be induced during the alignment process. Constitutional AI (Bai et al., 2022b), SELF-ALIGN (Sun et al., 2023d) and so on are more effective methods, where explicit value principles direct the training data construction or reward calculation. However, the target LLM has not yet directly learned to behave from these value principles. Actually, in-context learning is a promising method to prompt LLMs with clarified target value and regulate their behaviors (Ganguli et al., 2023). However, without fine-tuning, it cannot completely eliminate inherent harms. It is also challenging to express fine-grained value priorities and handle varying contexts via simple prompts. Therefore, future research should focus on developing efficient, stable alignment algorithms that transparently align LLMs with clear and generalizable target values instead of ambiguous proxies.

Automatic & Comprehensive Evaluation Accurate and robust benchmarks and evaluation methods are essential for guiding research about value alignment. At present, some benchmarks are constructed for alignment evaluation (Xu et al., 2023e; Sun et al., 2023a), which require human annotations or final human judgment. This makes them expensive and not easily scalable. Though powerful LLMs perform as an alternative for judgment, it highly relies on LLMs' capabilities and introduces uncertainty or biases. Consequently, automatic evaluation methods and metrics are urgently required to accelerate the assessment and research process. Evaluations across various abilities and difficulty levels should be considered: 1) understand and agree with human values; 2) diagnose scenarios involving values and make correct judgments; 3) perform consistently with human values, even in dilemmas; etc. This assessment becomes more and more difficult, from simple discrimination to exact behaviors, which attempts to detect the most essential values of LLMs behind their elicited behaviors. Since priorities among values can only matter in some quandary scenarios, we should also consider specific dilemma cases in the evaluation to figure out such fine-grained information.

7 Conclusion

This paper highlights the importance of specifying appropriate goals for big model alignment and presents the first survey of various alignment goals in existing literature. We propose a novel categorization for these goals in line with human learning process: human instructions, human preferences, value principles and basic values, which facilitates understanding their evolution paths and indicates further developments. To inspire studies aligning big models from the level of basic values, we discuss challenges and future directions. Besides, our survey provides a compilation of resources for big model alignment. We expect this survey to act as both a foundational guide and a source of inspiration for researchers and practitioners in this field.

8 Limitations

In this paper, we provide a comprehensive survey from the perspective of alignment goals for big models and present a novel categorization for these increasingly complex goals, which is in line with human learning hierarchy thus indicative for future research. Due to our emphasis on the evolution process of alignment goals, there may be some limitations in this paper.

Limited Details on Alignment Methods In terms of value alignment, there are two critical research questions: *what to align with?* and *how to align?* This study centers on the former one to clarify alignment goals, which performs as a premise for subsequent design of alignment methods. As a result, details about concrete alignment methods are not included in our paper, such as the reinforcement learning from human feedback (RLHF) and its improved versions. Information about these aspects is available in other surveys dedicated to LLMs alignment methodologies (Wang et al., 2023c; Zhang et al., 2023b), which differs from our paper in the reviewing perspective and are discussed by us in Appendix A.2.

Scope of Considered Big Models Examples of big models mainly include Large Language Models (LLMs) and Large Multimodal Models (LMMs). This survey and the taxonomy are primarily constructed on the alignment research of LLMs, and existing related works in the field of LMMs which still focus on the alignment goals of human instructions. As LMMs alignment develops, we argue that the proposed taxonomy should be applicable to LMMs as well. Besides, we would conduct future updates to include such advancement and ensure the comprehensiveness of our taxonomy.

9 Ethical Consideration

This paper conducts a comprehensive survey about alignment goals for big models, which aims at clarifying the most appropriate values encoded into AI and transparently guarantee their responsible development. Notably, discussing these details can also provide inspirations for designing malicious alignment goals, injecting harmful noises into the training data and adversarial attacks. More robust alignment methods are required at the same time.

References

2021. World values survey wave 7 (2017-2022).

2022. Pew global attitudes survey.

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*.

Umut Akcil, Huseyin Uzunboylu, and Elanur Kinik. 2021. Integration of technology to learning-teaching processes and google workspace tools: A literature review. *Sustainability*, 13(9):5018.

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*.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

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, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, Yan Xu, Samuel Cahyawijaya, Dan Su, Bryan Wilie, Romain Barraud, Elham J Barezi, et al. 2022. Towards answering open-ended ethical quandary questions. *arXiv preprint arXiv:2205.05989*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Duane Brown and R Kelly Crace. 2002. Life values inventory: Facilitator's guide. *Williamsburg, VA*.

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*.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023a. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023b. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

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

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*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

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*.

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*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Robert Gagne. The conditions of learning and theory of instruction robert gagné.

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*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Bernard Gert. 2004. *Common morality: Deciding what to do*. Oxford University Press.

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*.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Jesse Graham, Peter Meindl, Erica Beall, Kate M Johnson, and Li Zhang. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8:125–130.

Geyang Guo, Ranchi Zhao, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Beyond imitation: Leveraging fine-grained quality signals for alignment. *arXiv preprint arXiv:2311.04072*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020b. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Yue Huang, Qihui Zhang, Lichao Sun, et al. 2023a. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

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*.

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.

Richard Joyce. 2007. *The evolution of morality*. MIT press.

Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From values to opinions: Predicting human behaviors and stances using value-injected large language models. *arXiv preprint arXiv:2310.17857*.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

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.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. *arXiv preprint arXiv:2305.13735*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001*.

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*.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown, and William Yang Wang. 2022. Safetext: A benchmark for exploring physical safety in language models. *arXiv preprint arXiv:2210.10045*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. arxiv.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023a. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 3.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023c. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Ryan O Murphy, Kurt A Ackermann, and Michel JJ Handgraaf. 2011. Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. 2020. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 124–130.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

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.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11183–11191.

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*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Milton Rokeach. 1967. Rokeach value survey. *The nature of human values.*

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

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

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *CoRR*, abs/2303.16755.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

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.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *arXiv preprint arXiv:2309.00779*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

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

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023b. Moraldial: A framework to train and evaluate moral dialogue systems via moral discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2213–2230.

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. 2023c. 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. 2023d. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 650–662.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

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*.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. 2024. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023c. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Norbert Wiener. 1960. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410):1355–1358.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023a. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin

Jiang. 2023b. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023c. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

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

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. 2023e. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*.

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

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. *Blog post, May*, 3.

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, et al. 2023a. Chinese open instruction generalist: A preliminary release. *arXiv preprint arXiv:2304.07987*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023c. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

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

Zhaowei Zhang, Nian Liu, Siyuan Qi, Ceyao Zhang, Ziqi Rong, Yaodong Yang, and Shuguang Cui. 2023e. Heterogeneous value evaluation for large language models. *arXiv preprint arXiv:2305.17147*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

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*.

A Supplements of Introduction

A.1 Scope of References

To make the survey as comprehensive as possible, we review papers in recent years (mostly 2019-2024) from well-known conferences and journals, including the ACL, EMNLP, NAACL, NeurIPS, ICLR, arXiv where newly emergent papers are released, and so on. Topics of related work encompass LLMs alignment, value alignment, value evaluation, reward modeling, instruction tuning, etc.

A.2 Related Work

In this section, we review related work from two primary aspects: the surveys about AI alignment and the discussions on alignment goals.

With remarkable progress in big models, great efforts have been made to align them with human values and ensure their responsible development. To furnish a picture of existing works and inspire future research, there are numerous surveys about AI or large language model alignment. Zhang et al. (2023b) and Wang et al. (2023c) summarize research works about instruction tuning, including the available datasets, training methods, evaluation methods, applications to other modalities and domains. Shen et al. (2023) exhibit a more comprehensive survey of alignment methodologies by categorizing them into outer and inner alignment. Ji et al. (2023b) also explore the methodologies and practical applications of AI alignment. However, these studies predominantly explore the research question '*how to align*', focusing on the algorithms

Data Source	Dataset	#Tasks	#Instruction	Prompt Types
	PromptSource (Bach et al., 2022)	180	2,085	ZS
	P3 (Sanh et al., 2021)	270	2,073	ZS
	Natural Instructions (Mishra et al., 2021)	61	61	ZS & FS
	Super-NatInst (Wang et al., 2022b)	76	1,616	ZS & FS
Existing NLP Benchmarks	GLM-130B (Zeng et al., 2022)	74	-	FS
	xP3 (Muennighoff et al., 2022)	83	-	ZS
	OPT-IML Bench (Iyer et al., 2022)	1,991	18M	ZS & FS & CoT
	Flan 2022 Collection (Longpre et al., 2023)	1,836	15M	ZS & FS & Co
	COIG (Zhang et al., 2023a)	2k	200k	ZS
	Unnatural Inst (Honovich et al., 2022)	117	240k	ZS
	Self-Instruct (Wang et al., 2022a)	175	82k	ZS
	Alpaca (Taori et al., 2023)	175	52k	ZS & FS
Model-Generated	Baize (Xu et al., 2023c)	-	111.5k	Conversation
	UltraChat (Ding et al., 2023)	-	675k	Conversation
	Evol-Instruct (Xu et al., 2023b)	-	250k	Varying Complexity
	Phoenix (Chen et al., 2023b)	-	189k	Multilingual
	Bactrain-X (Li et al., 2023a)	-	3.4M	Multilingual
Crowd-Sourcing	ShareGPT (Chiang et al., 2023)	-	~100k	Converastion
	OpenAssistant (Köpf et al., 2023)	-	~161k	Conversation

Table 1: Details of public instruction datasets, ordered by their release time. ‘ZS’ and ‘FS’ mean zero-shot and few-shot respectively and ‘CoT’ means chain-of-thought.

rather than the underlying objectives. Differently, this paper provides an overview from a novel perspective of ‘*what to align with*’, which is critical to determine the objective encoded into AI.

In previous studies, there are a few discussions about defining precise and appropriate goals for alignment. For example, *Specification Problem* (Leike et al., 2018) underscores the necessity for precise reward modeling to ensure correct alignment. Furthermore, various alignment goals and their differences have been analyzed in position papers (Gabriel, 2020), ranging from instructions, intentions, preferences to interests and values. Distinguished from previous works, our paper conducts the first practical survey of alignment goals introduced in existing research works. By dissecting their essence and integrating the insights gained from human learning process, our paper presents a novel categorization with increasing abstraction and complexity. In addition, we also delve into the challenges and future research directions.

B Supplements of Human Instructions

Details of public instruction datasets are enumerated in Table 1.

B.1 Taxonomy of Alignment Goals

Figure 3 illustrates the taxonomy of alignment goals in our paper.

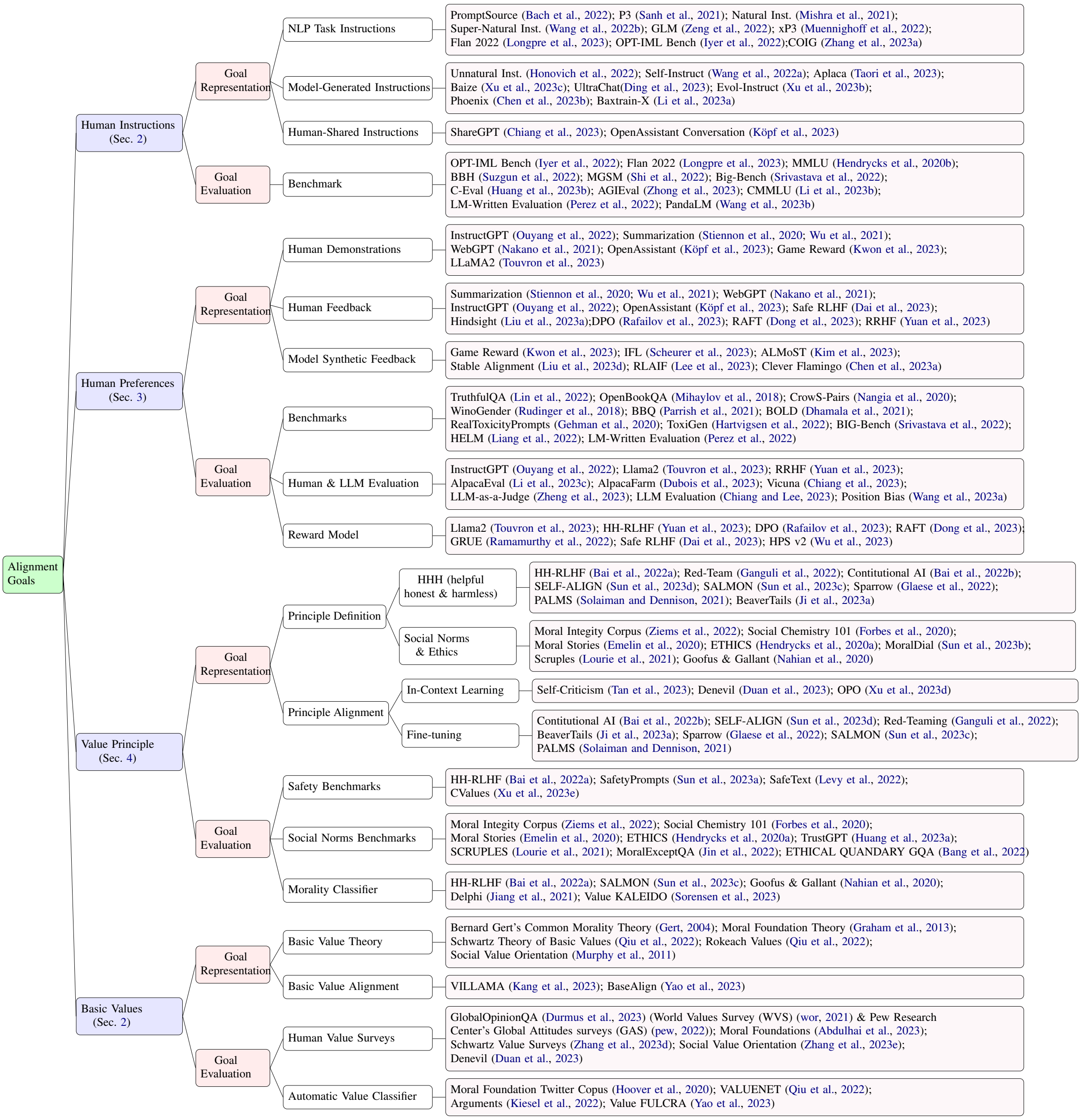

Figure 3: Taxonomy of reviewed papers about various alignment goals.