University of Catania

Mathematics and Computer Science

Luana Bulla

Natural Language Processing for Human-centered AI

Doctoral Thesis

Supervisor: Prof. Misael Mongiovi
Prof. Aldo Gangemi

Academic Year 2024–2025

Contents

Acknowledgements **2**

Introduction **3**

1 Human-Centered and Ethics AI **9**
1.1 Human-Centered AI: Definition and Framework 10
1.1.1 Foundational and Normative Definitions 11
1.1.2 Conceptual and Architectural Models 12
1.1.3 Sociotechnical Perspectives 14
1.1.4 Operational Frameworks and Design Guidelines 17
1.2 Ethical AI 19
1.2.1 Foundational Principles and Multi-Level Framework . 21
1.2.2 Contextual Morality 26
1.2.3 FAT: Fairness, Accountability and Transparency . . . 28
1.3 Human-Centered Approaches to Designing and Evaluating AI 38
1.3.1 Human-oriented Design Methods 38
1.3.2 Evaluation of AI 42
1.4 Humans Teaming with AI 45

2 Human-centered AI Methods: State of the art **48**
2.1 Moral Values Detection 49
2.1.1 Moral Foundation Theory 49
2.1.2 Basic Human Value 70
2.1.3 Morality-as-Cooperation 77
2.2 Towards Assessable and Controllable AI in Human-centered domain 79
2.2.1 Evaluating Subjective Labels: Annotator Diversity, Agreement, and Model Alignment 80
2.2.2 Grammar Constraints and Decoding Strategies 83
2.3 AI for the Deaf Community 86
2.3.1 Theoretical foundations of Sign Language translation . 87
2.3.2 Sign Language Translation: an Overview 89

3 Moral Value Detection **92**
3.1 Methodology . . . 95
3.1.1 LLM-based models . . . 96
3.1.2 NLI-RoBERTa-based model . . . 100
3.2 Results and Evaluation . . . 101
3.2.1 Dataset Preprocessing . . . 104
3.2.2 Comparison Among Language Models . . . 105
3.2.3 Human vs. Models Comparison . . . 118
3.3 Discussion . . . 124

4 A Novel Framework for Evaluating Classifiers in a Multi-perspective Domain **129**
4.1 A Unified Approach to Inter-Annotator Agreement and Evaluation . . . 131
4.1.1 Background . . . 132
4.1.2 A novel inter-annotator agreement metric based on F1-score . . . 133
4.2 Results and Evaluation . . . 137
4.2.1 Binary classification task . . . 137
4.2.2 Case study on moral values identification . . . 140

5 Grammar-Constrained Natural Language Generation **144**
5.1 A Novel Framework for Grammar- Constrained LLM Decoding 147
5.1.1 Background . . . 148
5.1.2 A Novel Expressive Grammar Notation for Constraining LLM Output . . . 150
5.1.3 Grammatically-Compliant Text Generation . . . 152
5.2 Results and Evaluation . . . 154
5.3 Limitations . . . 161

6 Leveraging Large Language Models for Accurate Sign Language Translation in Low-Resource Scenarios **165**
6.1 Methodology . . . 168
6.2 Experiments and Results . . . 171
6.2.1 Spoken-to-sign translation . . . 173
6.2.2 Sign-to-spoken translation . . . 176
6.3 Discussion . . . 178
6.4 Limitations . . . 179

7 Conclusion **181**

Abstract

Despite achieving human-level performance on established benchmarks, state-of-the-art Artificial Intelligence (AI) systems exhibit a drop in performance when applied to high-subjective, ethical tasks. These limitations highlight the need for a paradigmatic shift toward Human-Centered AI (HCAI), which prioritizes human values, contextual awareness, and societal benefit as core design principles.

This dissertation operationalizes the HCAI framework within Natural Language Processing (NLP) through integrated theoretical, methodological, and applied contributions. We first establish the conceptual and sociotechnical foundations of HCAI and conduct a comprehensive survey of the state of the art across three critical domains: ethical AI, robust and thrustworthy AI, and inclusive AI. Building upon this foundation, the dissertation addresses four interrelated research challenges: the alignment of language models with human moral values, the enhancement of controllability in generative systems, the development of novel evaluation frameworks for multi-perspective and highly subjective tasks, and the design of inclusive technologies for underrepresented communities, with a focus on sign language translation.

Collectively, these contributions advance NLP systems toward being more controllable, trustworthy, inclusive, and aligned with human values. This research supports the broader vision of HCAI by providing a pathway from a purely performance-oriented AI to one that is also responsible, interpretable, and genuinely human-centered, thereby advancing the field both theoretically and operationally.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Misael Mongiovì and Aldo Gangemi, for their guidance and support throughout these years of research. Their supervision has been fundamental to my scientific and personal development, and I am deeply grateful for the opportunities, insights, and trust they have provided along the way.

I also wish to thank all the members of the ISTC-CNR STLab group for their warm welcome and for the stimulating and collaborative environment that has always characterized our research meetings. My thanks also go to my colleagues at the ISTC-CNR group in Catania, whose daily collaboration and support have made my work both productive and enjoyable.

I am particularly grateful to Professor Carles Sierra and his research group at the IIIA-CSIC, for allowing me to broaden my research experience, explore new perspectives, and grow both professionally and personally during my stay there.

I would also like to thank the University of Catania for allowing me to undertake and complete this PhD journey.

Finally, I would like to express my heartfelt thanks to my family, for their constant encouragement, patience, and trust, which have been essential throughout this journey.

Introduction

Artificial Intelligence (AI) has recently undergone a significant evolution, advancing from early symbolic systems to the prevailing paradigms of machine learning and deep learning. Despite achieving performance comparable to human levels on established benchmarks, large-scale models still exhibit significant degradation when confronted with ambiguity, subjectivity, or out-of-distribution conditions [56, 305]. In contrast to humans, who flexibly adapt by leveraging contextual knowledge and social grounding, AI systems often lack robustness, raising critical concerns about their reliability in safety-sensitive applications. Moreover, the widespread adoption of technology-centered approaches in everyday tasks has highlighted pressing ethical and societal challenges, including well-documented cases of bias, opacity, and harmful consequences. These limitations highlight the urgent need for alternative paradigms that integrate human values, contextual awareness, and human oversight into both the design and evaluation of AI systems.

To address this challenge, the Human-Centered AI (HCAI) paradigm [117, 305, 320] represents a shift toward reframing AI research around human needs, capabilities, and fundamental rights. HCAI emphasizes the augmentation of human performance while promoting transparency, accountability, and ethical alignment. This dissertation contributes to the HCAI paradigm by operationalizing its principles in the context of Natural Language Pro-

cessing (NLP).

These challenges converge toward the general goal of building AI that is not only ethically aligned with human morality but also robust and controllable, in line with the HCAI principle of trustworthiness. Furthermore, I propose a novel validation framework designed to assess whether alignment with ethical AI can be meaningfully evaluated in multi-perspective, highly subjective scenarios. Finally, this dissertation expands the boundaries of current AI systems by contributing to more inclusive technologies, with a particular focus on supporting the Deaf community, thereby advancing the inclusivity principle at the core of HCAI.

This dissertation's principal contributions are:

- I present a comprehensive survey of the theoretical foundations of the HCAI paradigm, reviewing its normative definitions, conceptual and architectural models of human–AI interaction, and sociotechnical perspectives that position AI within institutional and societal ecosystems. I further analyze operational frameworks that translate principles into design guidelines, systematically discussing the core principles of Ethical AI (i.e. Fairness, Accountability, and Transparency) together with related notions of autonomy, beneficence, and non-maleficence. The review exposes both convergences and critical gaps, including the risk of "principle-washing" and the need for context-sensitive approaches that account for cultural pluralism in moral reasoning.

- I present a comprehensive survey of the state of the art in HCAI, focusing on three central domains: (i) moral value detection, detailing main theoretical taxonomies, supervised and zero-shot approaches, and the challenges of subjectivity in annotation; (ii) the controllability of AI systems, focusing on inter-annotator agreement in subjective tasks,

and grammar-constrained strategies for LLM generation; and (iii) AI for inclusivity, highlighting advances and limitations in sign language technologies and accessibility-oriented NLP.

- In the context of moral value detection, I examine the performance of a range of LLMs — including GPT-4, GPT-3.5, LLaMA-70B, Mixtral-7x8, Mistral-7B, and LLaMA-7B — across different parameter scales. I further assess the effectiveness of prompt design by testing the models in two distinct prompting scenarios, analyzing how prompt structures influence performance across varying levels of task difficulty. To obtain a nuanced evaluation, I separately measure performance on multi-label moral value detection (i.e., identifying specific values) and binary moral classification (i.e., detecting the presence or absence of moral content). Finally, I introduce a direct model–human comparison framework for moral detection, designed to mitigate the subjectivity inherent in existing gold standards.

- I introduce a direct model–human comparison framework for evaluating classification tasks, designed to address the subjectivity limitations of gold standards and extend the applicability of traditional machine learning metrics such as precision, recall, and F1 score. As part of this framework, we propose a novel Inter-Annotator Agreement metric, F1-kappa, which generalizes F1 to more than two annotators and normalizes by the expected value analogously to Fleiss' Kappa. F1-kappa is broadly applicable to categorical data and enables direct comparison between human annotations and model performance. I conduct a comprehensive evaluation of the proposed metric across binary, multi-class, and multi-label scenarios, demonstrating its robustness and versatility

under diverse conditions. Finally, through a case study on a multi-label moral value dataset, I show that the framework reveals modern LLMs achieving parity with human performance.

- In the domain of controllable AI, I present GRAMMAR-LLM, a framework that integrates formal grammars into the generation process of language models. This approach enforces strict adherence to predefined syntactic and structural requirements, overcoming the limitations of existing methods in fine-grained text control. GRAMMAR-LLM is based on a novel formalization that generalizes the LL(1) class of Context Free Grammars (CFGs), allowing users to define grammars without relying on tokenizer-specific details, thereby enhancing expressiveness while preserving efficiency. GRAMMAR-LLM demonstrates adaptability across different tasks — including hierarchical classification, sign language translation, and semantic parsing — and across multiple model architectures (LLaMA-3, AMRBART). Beyond enabling controllability, experimental results also show consistent performance improvements in zero-shot, one-shot, few-shot, and fine-tuned settings, underscoring the broad applicability of the framework.

- In the area of AI for inclusivity, I present AulSign, a novel sign language translation framework that leverages LLMs to translate to and from languages not seen during training. Designed for low-resource environments, AulSign addresses data scarcity by integrating external lexicons and structured linguistic representations into a unified pipeline. The model maps signs to a meta-lexicon, providing explainability in the translation process by enabling users to trace inference steps, detect misalignments, and interpret translation errors. Transla-

tion accuracy and usability are further enhanced through intermediate representations such as SignWriting and SMPL-X, which support the generation of visual content (e.g., animated avatars) and improve accessibility in spoken-to-sign applications. Through comparative experiments across varying data availability scenarios and multilingual settings, AulSign consistently outperforms state-of-the-art models in both spoken-to-sign and sign-to-spoken translation tasks.

The dissertation is structured as follows. Chapter 1 establishes the theoretical foundations of the HCAI paradigm, outlining its definitions, conceptual frameworks, sociotechnical perspectives, and the ethical principles of Fairness, Accountability, and Transparency (FAT). Building on this foundation, Chapter 2 surveys the state of the art in HCAI, with a focus on three domains that frame the core contributions of this work: moral value detection, the challenges of developing assessable and controllable AI, and the emerging field of AI for inclusivity. The subsequent chapters develop these themes into specific methodological and experimental contributions. Chapter 3 develops the first contribution by providing an extensive assessment of moral value detection models, analyzing the impact of model size, training data, and prompt design, and introducing a direct model–human comparison framework. Chapter 4 introduces a novel inter-annotator agreement metric (F1-kappa) able to compare AI models with human performances in highly subjective settings, while Chapter 5 details a novel framework for grammar-constrained text generation. Chapter 6 proposes AulSign, a methodology for low-resource sign language translation. Collectively, these contributions aim to advance NLP systems that are not only technically proficient but also more robust, inclusive, and aligned with human values and well-being. This promotes the transition from a performance-oriented AI to one that

is also responsible, interpretable, and genuinely human-centered, thereby advancing the HCAI paradigm both theoretically and operationally.

1 Human-Centered and Ethics AI

The evolution of AI from its symbolic and rationalist foundations to contemporary machine learning (ML) and deep learning paradigms highlights both the field's advances and its limitations. Although current models' results are comparable or even superior w.r.t. human performance, they present a drop in performance when faced with ambiguity, novelty, or adversarial conditions. In contrast to humans, who flexibly adapt by leveraging contextual knowledge and experience, AI models often fail to demonstrate robustness in out-of-distribution settings, raising critical concerns for their reliable deployment [320, 361].

Beyond technical challenges, the diffusion of technology-centered AI systems has given rise to significant ethical and societal concerns. Well documented cases (e.g., the COMPAS risk assessment tool, which exhibited racial bias against African American defendants [117], and Google Flu Trends, which failed to adapt to shifts in user behavior [320]) demonstrate the possibilities in AI systems of perpetuating bias, lacking transparency, and generating harmful outcomes in high-stakes domains. These cases highlight the need to introduce values, cognitive constrains and social contexts in AI model design [361].

In response, the HCAI paradigm has been introduced to reorient AI research and practice around human needs, capabilities, and rights. Following

Shneiderman et al. [320] and Xu et al. [361], the HCAI paradigm emphasizes the development of systems that are transparent, accountable, and ethically aligned—designed not to replace humans, but to augment human performance and well-being.

This chapter aims to outline the theoretical and methodological foundations of the HCAI paradigm. Section 1.1 examines definitions and conceptual frameworks for HCAI, while Section 1.2 addresses the domain of Ethical AI, including normative principles, multi-level frameworks, and the Ethical AI principles of fairness, accountability, and transparency. Section 1.3 focuses on human-centered approaches to design and evaluation, with particular attention to participatory and value-sensitive methodologies. Finally, Section 1.4 discusses human–AI teaming, exploring cooperative models in which AI functions as a collaborator rather than as a substitute.

Together, these perspectives establish the conceptual and operational foundation for the remainder of the dissertation. They demonstrate that integrating ethical, social, and design considerations is essential for developing AI systems that are robust, inclusive, and aligned with human well-being.

1.1 Human-Centered AI: Definition and Framework

The field of HCAI integrates a wide spectrum of perspectives, spanning four key areas: foundational and normative definitions; conceptual and architectural models; sociotechnical analyses; and operational frameworks with associated design guidelines. Foundational and normative definitions (cf. Sect. 1.1.1) establish HCAI as an ethically grounded paradigm that prioritizes human values and agency. Conceptual and architectural models (cf.

Sect. 1.1.2) offer methodologies for designing systems that meaningfully integrate human oversight and control. Sociotechnical perspectives (cf. Sect. 1.1.3) examine AI's embeddedness within complex societal and institutional ecosystems, emphasizing contextual factors that shape technology adoption and impact. Operational frameworks (cf. Sect. 1.1.4) bridge theory and practice by translating abstract principles into concrete design methodologies and evaluation criteria. Collectively, these areas explore the alignment of AI technologies with human values, capabilities, and societal needs.

1.1.1 Foundational and Normative Definitions

To establish the theoretical groundwork for HCAI, research and institutions frame it as a normative paradigm grounded in ethical, civic, and societal imperatives.

Specifically, Shneiderman et al. [319] define HCAI as a design paradigm focused on augmenting and supporting human performance in trustworthy, safe, and empowering ways, thereby reinforcing self-efficacy, creativity, and civic participation. This vision is later extended by embedding Human-Centered Design (HCD) principles throughout the entire AI lifecycle [322]. While Shneiderman et al. emphasize design empowerment, Holzinger et al. [146, 147] advocate for a broader interdisciplinary convergence. They conceptualize HCAI as a synergy between natural and artificial intelligence, necessitating a strong integration with legal, ethical, and safety standards.

Building upon these academic foundations, policy-oriented institutions operationalize the normative framework of the HCAI into concrete guidelines and visions. The Stanford Institute for HCAI [328] frames this field as a vision oriented toward enhancing well-being, public good, and societal trust by drawing inspiration from human behavior. Building on this human-centric

foundation, the European Commission's AI High-Level Expert Group [326] provides a policy-oriented definition that is explicitly grounded in the moral status of human beings, aiming to prioritize fundamental rights.

1.1.2 Conceptual and Architectural Models

Moving from normative foundations to concrete system design, several studies propose architectural frameworks for conceptualizing HCAI systems [269, 305, 359, 361]. While these frameworks differ in their granularity and focal concerns, they share the common aim of operationalizing HCAI principles throughout both the design and implementation phases.

Specifically, Xu et al. [359, 361] and Schmager et al [305] present two different interaction-focused frameworks. Xu et al. focus on interactional implications through three interdependent pillars: human factors engineering, human-reflective technology, and ethically aligned design (cf. Figure 1.1). Human factors engineering ensures system usability and explainability while preserving human agency in critical decision-making processes. Human-reflective technology structures AI to augment human capabilities within socio-technical ecosystems, emphasizing human-machine complementarity. Ethically aligned design systematically integrates core ethical principles (i.e. fairness, justice, privacy, and accountability) to ensure meaningful human control and decision traceability. By emphasizing interaction, this perspective positions human–AI engagement as the core process through which HCAI principles are enacted in practice.

Building on this conceptualization, Schmager et al. propose a framework that focuses on augmenting human capabilities and safeguarding meaningful human control, while also stressing context-sensitivity, compliance with ethical and legal standards, and the promotion of social well-being. This

contribution reinforces the interactional perspective by grounding HCAI in established human-centered design methodologies.

In contrast to these interaction-focused approaches, Serafini et al. [314] and Pyae et al [269] present more empirically-driven frameworks.

Specifically, Serafini et al. provide system-level abstractions by decomposing HCAI into four interconnected core components: observations (environmental data collection), requirements (human-specified tasks), actions (agent-executed operations and their effects), and explanations (behavioral justifications that foster trust and accountability). This contribution goes beyond architectural conceptualization by introducing a taxonomic classification that differentiates HCAI models into logical, probabilistic, functional, and hybrid types (cf. Figure 1.1). This taxonomy systematically maps technical approaches to their corresponding reasoning paradigms, providing a comparative schema for evaluating HCAI implementations at the system level.

Pyae [269] organizes HCAI principles into empirically derived hierarchies, encompassing 26 attributes across four stratified levels: ethical foundations, usability, cognitive and emotional dimensions, and personalization (Figure 1.1). This framework positions ethics as the foundational layer, embedding values such as fairness, transparency, privacy, trust, and safety as core prerequisites for human-centered AI. The usability layer emphasizes human-centric design principles and operational effectiveness, ensuring systems remain intuitive, efficient, and supportive of human autonomy and control. Cognitive and emotional dimensions address sophisticated aspects of human-AI alignment by integrating empathy, well-being, and user experience considerations into system design. Personalization constitutes the top layer, highlighting adaptive mechanisms that calibrate AI behavior to

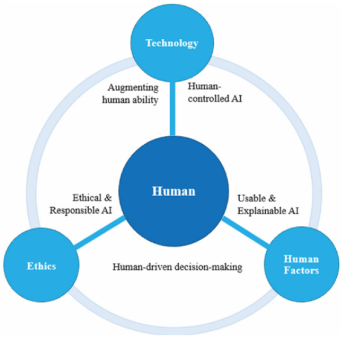
individual preferences, goals, and contextual requirements. Empirically validated through a mixed-method study involving 120 AI professionals, this hierarchical organization provides a systematic foundation for assessing and prioritizing attributes essential to HCAI realization across different levels of implementation.

1.1.3 Sociotechnical Perspectives

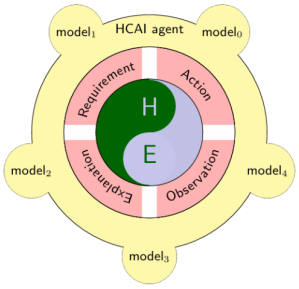
While HCAI frameworks such as those of Xu et al. [361] and Pyae et al. [269] emphasize system design and architecture at a conceptual level, a growing consensus recognizes AI as an inherently sociotechnical construct, deeply embedded in human and institutional contexts. This perspective promotes frameworks that move beyond purely AI technical system views, to foreground the broader social and institutional environments in which AI systems operate. These contributions advance a more integrated and holistic account of HCAI, diverging in methodological orientation: some focus on micro-level interaction dynamics, while others articulate broader design paradigms or address systemic issues of governance and oversight [29, 91, 117, 285].

At the micro-level of user interaction, Riedl et al. [285] and Auernhammer [29] theorize cross-disciplinary frameworks that bridge user experience and machine learning, while promoting advanced and ethically responsible AI systems. Specifically, Riedl et al. highlight the need for systems that can both understand sociocultural contexts and offer interpretable, user-accessible explanations, placing emphasis on transparency, fairness, and user understanding. In this context, Auernhammer [29] proposes a comprehensive HCAI design approach that integrates three complementary perspectives: rationalistic, humanistic, and judicial (cf. Figure 1.2). The ratio-

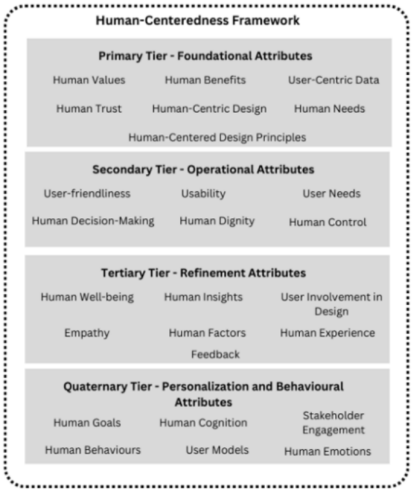
Conceptual and Architectural Models



The HCAI framework with specified design goals according to Xu et al. (2022)



Schematic representation of an HCAI agent, according to Serafini et al. The external ring denotes the agent, which interacts with humans (H) and the environment (E) in the inner ring through the components of the middle ring, supported by one or more internal models



Pyae et al. Human-Centeredness Framework

Figure 1.1: Xu et al. [361], Serafini et al. [314] and Pyae et al. [269] HCAI frameworks

nalistic lens focuses on the technological dimension, encompassing mathematical and computational advancements aimed at replicating human abilities while also providing general principles for ethical, lawful, and robust AI. The humanistic perspective emphasizes AI as a problem-solving tool to enhance human capabilities and improve human conditions, grounded in human-centered design practices that account for embodiment, interaction, and context-specific ethical challenges often overlooked by abstract principles. Finally, the judicial perspective highlights the need for adaptive legal frameworks and policies that co-evolve with technological innovation and address the broader societal implications of AI.

At the macro level, Garibay et al. [117] extend the discussion to systemic governance and institutional oversight, articulating six grand challenges for HCAI: human well-being, responsible design, privacy, user-centered interaction, independent oversight, and cognitive alignment (Figure1.2). The first emphasizes the design of AI systems that actively foster human flourishing and collective well-being, while accounting for risks introduced by automation and biased training data. Responsible design encompasses legal, ethical, and moral considerations, promoting principles such as explainability, fairness, accountability, and reliability to counteract the increasing opacity of responsibility in advanced machine learning. Privacy highlights the importance of safeguarding personal data by ensuring informed consent, establishing clear limitations on data usage and retention, and protecting against misuse or unauthorized access. User-centered interaction stresses the integration of human-centered design and evaluation frameworks throughout the AI lifecycle, ensuring technologies are developed with and for diverse users, empowering human performance while preserving control. Independent oversight addresses the governance dimension, calling for mechanisms

grounded in Environmental, Social, and Governance (ESG) principles as well as guiding values such as fairness, integrity, resilience, and explainability, thereby fostering trust and accountability. Finally, cognitive alignment focuses on the Human–AI interaction, ensuring that AI systems respect cognitive processes, support human competencies, and strike a balance between autonomy and human oversight in work and everyday contexts.

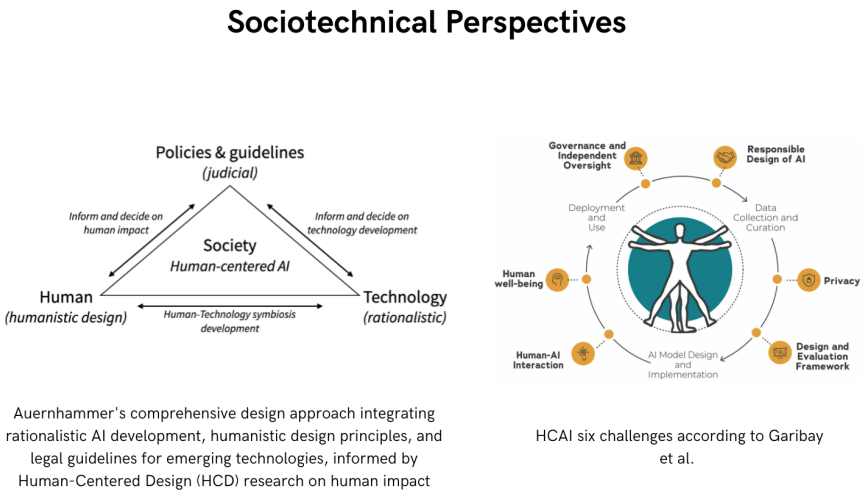


Figure 1.2: Auernhammer [29] and Garibay et al. [117] HCAI frameworks

1.1.4 Operational Frameworks and Design Guidelines

Building on these theoretical foundations and sociotechnical insights, Xu and Gao [360] and Amershi et al. [17] propose more pragmatic frameworks that operationalize HCAI principles into engineering practice.

Specifically, Xue and Gao address systemic implementation through a comprehensive multi-level framework that systematically links seven human-centered design goals to 28 design principles and 15 implementation approaches. These components provide specific and actionable guidance for

project teams and organizations, with design principles explicitly mapped to overarching goals and implementation approaches serving as tactical "how-to" methodologies. Structured across user, technology, and ethics dimensions that establish synergistic integration of humans and intelligent systems as organically unified human-machine ecosystems, the framework promotes interdisciplinary practices while aligning interventions across project and organizational levels (cf. Figure 1.3). Furthermore, this framework advances the design of systems through an integrated and well-defined HCAI process that synthesizes human-centered design approaches (exemplified by the widely adopted "double diamond" methodology) with conventional AI lifecycle management and established HCAI guidance protocols. To promote systematic adoption, the framework proposed by Xue and Gao delineates a "three-tiered" implementation strategy: at the macro-societal level, cultivating interdisciplinary talent, establishing governmental funding mechanisms, and developing HCAI standards; at the organizational level, fostering an HCAI-oriented culture, formulating institutional guidelines, and standardizing processes; and at the AI project team level, assembling multidisciplinary teams to operationalize HCAI processes in practice.

Amershi et al. operationalize their framework by distilling 150 design recommendations into 18 empirically validated guidelines for human–AI interaction. These guidelines are structured along the temporal dimension of user engagement (from initial system encounter to long-term adaptation) and address key themes such as intelligibility, error handling, user feedback, and transparency. Representative examples include making system capabilities explicit, supporting efficient error correction, and clearly conveying the consequences of user actions. Validated through extensive user studies and expert review, these guidelines provide an empirical foundation for

translating ethical principles into concrete interaction behaviors, rendering them particularly salient for HCI practitioners seeking actionable support in interface design.

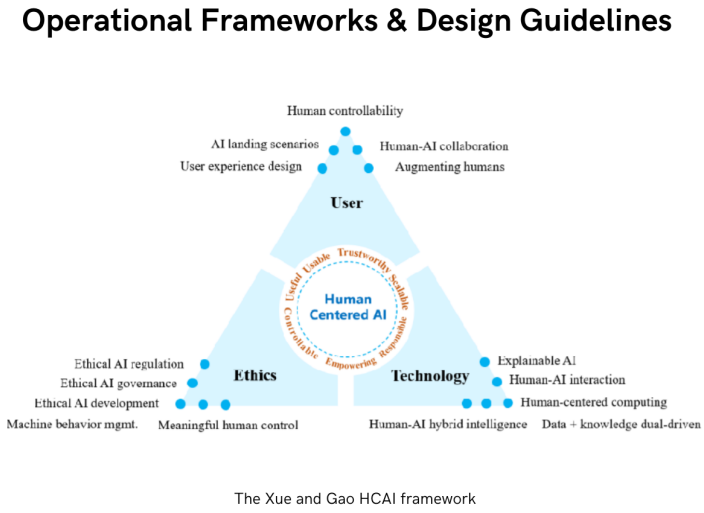


Figure 1.3: Xue and Gao [360] HCAI frameworks

1.2 Ethical AI

Ethical AI has become a critical area of research in response to the increasing deployment of AI systems in high-stakes domains where algorithmic decisions may significantly affect human lives and societal structures [167, 293]. This research field aims to integrate human values and fundamental rights into AI development, advocating transparency, accountability, fairness, and privacy as indispensable requirements rather than optional considerations [175]. A growing body of scholarship and policy work has recently articulated ethical frameworks and guidelines, focusing on the importance of ongoing human involvement throughout the AI lifecycle, from annotation and verification to governance [9, 15, 40, 48, 50, 51, 60, 63, 65, 67, 68, 69,

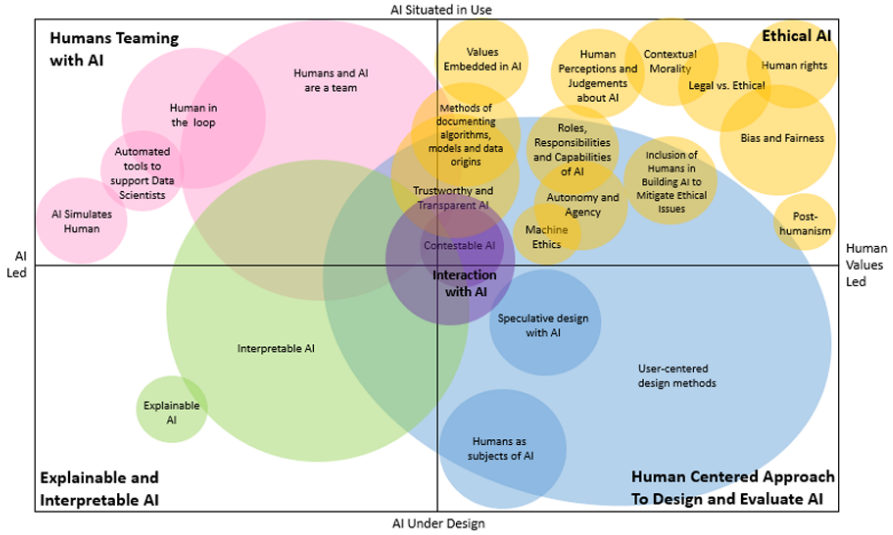


Figure 1.4: HCAI scheme [65]

70, 72, 76, 77, 83, 92, 94, 101, 103, 112, 119, 138, 140, 142, 145, 152, 180, 182, 189, 194, 195, 197, 200, 202, 207, 215, 217, 221, 222, 225, 232, 233, 235, 236, 244, 250, 253, 256, 259, 268, 281, 282, 283, 286, 287, 288, 296, 297, 298, 301, 303, 309, 316, 317, 319, 320, 321, 347, 350]. Within the HCAI paradigm, ethics assumes a transversal role, permeating all stages of design and deployment to ensure that AI systems function not only with technical robustness but also in alignment with broader social values.

Section 1.2.1 presents an overview of the Ethical AI foundational principles and multi-level frameworks. In Section 1.2.2, we examine the challenges of moral pluralism in aligning AI systems with different cultural and social values, while Section 1.2.3 focuses on the Fairness, Accountability, and Transparency (FAT) principles, which are the three pillars at the core of scholarly and policy discussions on ethical AI.

1.2.1 Foundational Principles and Multi-Level Framework

Current approaches in AI ethics design foundational principles that constitute the normative framework for responsible AI development. Early large-scale mappings of ethical guidelines have identified recurring clusters of values.

Jobin et al. [167] provide a comprehensive analysis of 84 ethical guidelines, identifying five core principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. Transparency concerns making AI decision-making processes intelligible; justice and fairness address equitable treatment and bias mitigation; non-maleficence refers to the duty to avoid harm; responsibility entails clear accountability for AI outcomes; and privacy safeguards personal data and individual rights. Fjeld et al. [109] similarly mapped 36 sets of principles and identified eight recurrent themes, including transparency, fairness, privacy, accountability, safety, professional responsibility, human oversight, and promotion of human values. Hagendorff [133] provided a critical evaluation of 22 guidelines, highlighting both convergence (e.g., around transparency, fairness, and accountability) and notable omissions, such as sustainability. Following Jobin et al., Khan et al. [175] propose a systematic literature review that identifies 22 ethical principles, with transparency (mentioned in 17 studies), privacy (16 studies), accountability (15 studies), and fairness (14 studies) emerging as the most frequently cited concerns.

Building on these foundational contributions, Floridi [111] advances a unified framework consisting of five key principles for AI in society: beneficence (promoting well-being), non-maleficence (avoiding harm), autonomy (safeguarding human agency), justice (ensuring fairness), and explicability (enabling understanding). In contrast to Jobin et al., Floridi explicitly in-

troduces beneficence and autonomy. Whereas non-maleficence establishes a negative duty to avoid harm, beneficence articulates a positive obligation to promote human well-being and generate tangible societal benefits. Autonomy underscores the preservation of meaningful human control over decisions and actions, ensuring that critical choices remain subject to human oversight, particularly in ethically and legally significant domains such as healthcare, justice, and politics. Furthermore, Floridi reconceptualizes transparency through the broader principle of explicability, which encompasses not only openness but also the traceability and intelligibility of AI decisions. Within this framework, responsibility and privacy are not treated as standalone principles but are instead subsumed under explicability.

From a policy perspective, institutional actors have progressively developed ethical frameworks that seek to operationalize normative principles and inform regulatory and governance practices. The European Commission's Ethics Guidelines for Trustworthy AI [7] define seven key requirements: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and accountability. In contrast to the preceding academic frameworks, which primarily formulate abstract normative principles, the European Commission's guidelines adopt a more operational and policy-oriented orientation. They reconceptualize ethical values as concrete, actionable requirements explicitly designed to inform governance structures, regulatory frameworks, and implementation practices throughout the AI lifecycle. Internationally, the OECD AI Principles[1] and the UNESCO Recommendation on the Ethics of AI[2] provide intergovernmental soft-law instruments that have shaped national strategies, while the Council of Europe's

[1]https://www.oecd.org/en/topics/ai-principles.html
[2]https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

Framework Convention on AI [3] represents the first binding treaty on AI and human rights. Complementing these, the EU AI Act [3] establishes a comprehensive risk-based regulatory framework, marking a transition from ethical principles to enforceable obligations.

Complementing normative and policy-oriented approaches, technical organizations have advanced standards and operational frameworks to support the practical implementation of ethical principles. The U.S. NIST released its AI Risk Management Framework [334], structured around the functions of Govern, Map, Measure, and Manage, with a subsequent generative AI profile. The IEEE contributes with Ethically Aligned Design and the 7000-series standards, which cover issues ranging from transparency to well-being metrics (IEEE 7010) [304].

At the same time, critical perspectives emphasize the limits of principles when formulated in isolation. Mittelstadt et al. [234] argue that principles, when considered in isolation, are insufficient without clearly defined implementation pathways. Similarly, Winfield et al. [355] highlight that ethical governance must be embedded throughout the AI lifecycle, rather than treated as an add-on or post hoc consideration. These perspectives reframe the goal of defining ethical AI principles to ensuring their effective operationalization within concrete contexts and application domains. Greene et al. [130] similarly warn against the depoliticization of ethics through principle-washing, and Birhane [45] stresses the importance of situated, plural, and decolonial approaches. This critique has led to the development of multi-level frameworks that structure ethical considerations across different scales of impact, capturing the complexity and interconnectedness of ethical challenges. Within this perspective, ethical frameworks often advo-

[3] https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

cate value-sensitive design strategies that incorporate socio-cultural awareness, acknowledging that communities may prioritize values in different and sometimes conflicting ways [72, 189, 197, 281, 303] (cf. Sect. 1.2.2).

Current ethical multi-level frameworks target different dimensions, spanning stakeholder engagement [355], process design [294], and domain-specific concerns. In this setting, Huang et al. [156], Ryan et al. [294], Lu et al. [214] and Wang and Blok [349] represent significant contributions, providing different ethics-oriented multi-level frameworks. Specifically, Huang et al. propose a framework that groups AI-related ethical issues across three interconnected levels: individual, societal, and environmental. At the individual level, AI systems influence personal safety, privacy, autonomy, and human dignity (cf. Figure 1.5). The societal level addresses broader concerns, including fairness and justice, responsibility and accountability, transparency, surveillance and datafication, and democratic governance. Finally, the environmental level reflects the growing awareness of AI's ecological footprint, encompassing resource consumption, energy costs, and sustainability. In particular, the carbon emissions associated with large-scale AI training and deployment represent a critical consideration for responsible AI development [156]. Mäntymäki et al. [223] introduce the Hourglass Model, which connects environmental, organizational, and technical levels of governance. Lu et al. [214] propose a pattern catalogue to operationalize responsible AI in multi-level governance and engineering contexts. Wang and Blok [349] further extend this approach by distinguishing artifact-level and structural-level challenges, highlighting issues often overlooked by frameworks focused solely on technical artifacts.

In summary, existing approaches to ethical AI range from the formulation of high-level normative principles to their translation into operational

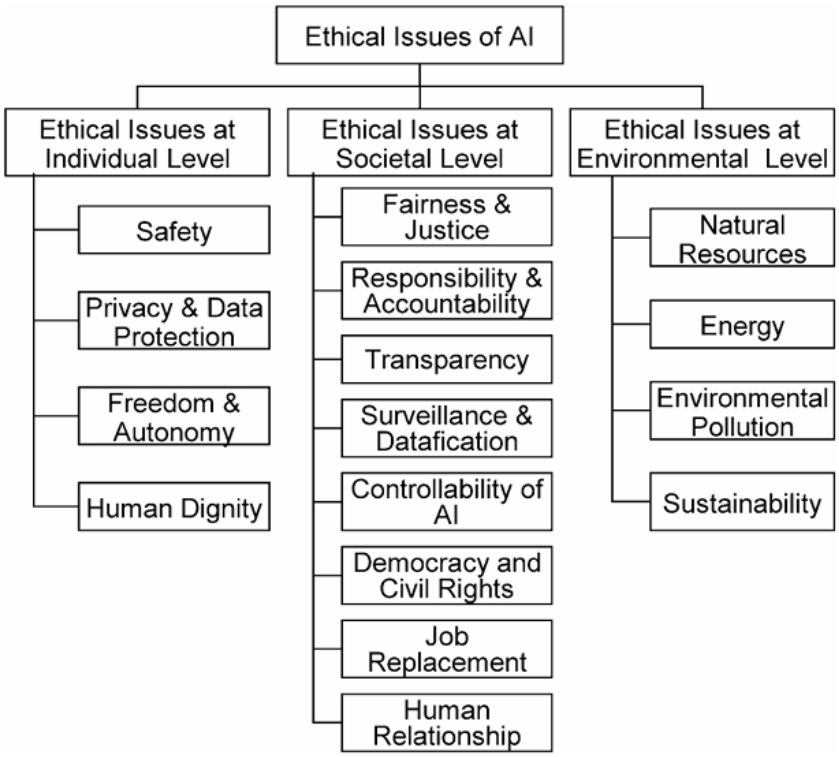


Figure 1.5: Multi-level framework of ethical AI [156]. The framework groups AI-related ethical issues into three interdependent levels: individual, societal, and environmental

requirements and to multi-level frameworks that integrate individual, societal, and environmental dimensions. Despite differences in scope and orientation, a recurring convergence emerges around key concerns such as transparency and explicability, justice and fairness, responsibility, and privacy. At the same time, critical perspectives highlight the necessity of embedding ethics throughout the AI lifecycle rather than treating principles as stand-alone declarations.

1.2.2 Contextual Morality

The previous section explored foundational principles and multi-level frameworks that articulate a normative baseline for HCAI. Despite attempts to establish a more coherent theoretical foundation, the implementation of these frameworks remains challenging and contested. A main challenge lies in accommodating moral pluralism, namely the recognition that ethical values and normative expectations vary across cultural traditions, social configurations, and situational contingencies [282, 321]. Such heterogeneity complicates the assumption that abstract, universally formulated principles can be consistently applied across diverse contexts, thereby raising fundamental questions about the cultural framing of ethical alignment in AI.

The NLP field engages with the operationalization of contextual morality, exploring a variety of tasks and methodological approaches. Foremost among these efforts is the task of moral or value classification, whereby textual data are systematically aligned with predefined normative taxonomies, such as those derived from moral psychology (cf. Chapter 3). While such taxonomies provide a useful abstraction and a starting point for systematizing moral concepts, they inevitably achieve generalization at the expense of both explainability and granularity. By reducing complex ethical reasoning to broad categorical labels, they risk obscuring the nuanced and context-dependent dimensions through which moral judgments are constructed. This limitation emerges not only at the cultural level, where underrepresented traditions are often excluded, but also at the individual level, where a person's moral stance is shaped by multiple, and sometimes conflicting, values derived from their unique cultural and biographical background. In this respect, moral taxonomies cannot fully capture the multifaceted and situated character of ethical evaluation, thereby exposing the tension between the as-

piration to universalize computational models and the inherently contextual nature of moral reasoning. To address these challenges, the HCAI paradigm has advanced approaches explicitly oriented toward human-centered design (cf. Section 1.3.1), with the aim of preserving the finer granularity of individual moral perspectives.

A further critical dimension concerns the alignment of AI systems with underrepresented cultures. Training datasets frequently privilege dominant cultural and linguistic groups, thereby marginalizing perspectives from less represented communities. As a consequence, AI models often reproduce a "default morality" reflective of the data on which they are trained, rather than aligning with more nuanced or localized orientations. Efforts to mitigate this problem involve not only curating more representative datasets, but also designing data collection processes that incorporate human oversight and active participation. Human-in-the-loop strategies (cf. Section 1.4) focus on the centrality of human judgment in annotating, validating, and refining training data, thereby reducing the risk of embedding unexamined biases. Similarly, methodologies emerging within the field of Ethical AI research on fairness (cf. Section 1.2.3) underscore the importance of collaborative practices in dataset construction, where human actors contribute not only to labeling but also to the conceptualization of categories and the exclusion of distortions. These practices underscore that the representativeness of datasets cannot be reduced to a purely technical concern; rather, it demands ongoing human engagement to foster alignment between computational systems and heterogeneous cultural and moral standpoints.

Ultimately, engaging with contextual morality compels a re-examination of prevailing evaluation practices and their adequacy for capturing context-sensitive ethical dimensions. Prevailing evaluation metrics generally rely on

aggregate performance measures over heterogeneous datasets, a practice that effaces cultural specificities and fails to account for divergences in system behavior across diverse moral communities. To overcome these limitations, alternative evaluation protocols are being developed within Human-AI design (cf. Section 1.3.2) and evaluation methodologies (cf. Chapter 4) , with the aim of achieving more fine-grained and context-sensitive assessments. Such approaches ensure that AI systems are evaluated not against abstract averages but in relation to the plurality of moral environments in which they operate.

These perspectives highlight the transversal character of contextual morality. Addressing this challenge requires methods that are not only technically robust but also interpretable and explainable (cf. Sect. 1.2.3). Transparency in the processes of moral reasoning provides a basis for scrutiny across heterogeneous cultural standpoints, whereas representative evaluation protocols secure validation of AI systems against the diverse moral norms that structure real-world contexts. In this respect, contextual morality operates across technical, methodological, and normative registers, while simultaneously engaging with wider discourses on fairness, accountability, and transparency in Ethical AI.

1.2.3 FAT: Fairness, Accountability and Transparency

Fairness, Accountability, and Transparency (FAT) constitute three foundational principles in the design and evaluation of human-centered and ethical artificial intelligence systems. Fairness seeks to ensure that algorithmic decisions do not systematically disadvantage individuals or groups on the basis of sensitive attributes such as gender, ethnicity, or socioeconomic status [65, 117]. Such biases arise from historically skewed datasets or from al-

gorithms inattentive to human values [61], producing harmful consequences in healthcare, justice, and social media [117]. The principle of accountability presupposes transparent allocation of responsibility and systematic traceability in decision-making processes, thus permitting stakeholders to scrutinize, evaluate, and contest algorithmic outcomes [258]. Transparency emphasizes interpretability and openness concerning model behavior, data provenance, and design choices, thereby fostering trust among developers, regulators, and end users [80, 335].

Together, these principles establish a mutually reinforcing framework in which fairness ensures equitable outcomes, explainability fosters understanding and trust, and accountability provides mechanisms of responsibility together with avenues for redress. However, while closely related, these principles constitute distinct yet complementary dimensions of ethical AI. Fairness primarily concerns the outcomes of AI systems, ensuring that decisions are free from unjustified bias and that individuals and groups are treated equitably. In contrast, Trasparency focuses on the process underlying decision-making, offering tools to understand how and why a model produces a given output. Accountability extends further by establishing responsibility for a system's behavior and consequences, while ensuring the presence of mechanisms for remediation in cases of error, harm, or misuse. Given their correlated but distinct character, these dimensions need not cooccur. For instance, a system may exhibit interpretability while producing unjust outcomes, or achieve fairness in outputs without affording transparency into its decision-making processes. Similarly, fairness and explainability may coexist in the absence of a clear framework of accountability.

Next, we examine these three ethical principles as applied to NLP: bias and fairness, transparency and explanability, and accountability.

Bias and Fairness

In the field of NLP, ensuring that AI models adhere to principles of fairness has recently become a crucial concern, encompassing tasks such as bias detection, analysis, and mitigation [46, 118, 277, 331]. This issue has become particularly relevant with the advent of LLMs, which often replicate and, in some cases, amplify pre-existing social biases embedded in their training data [115, 136, 262]. Unmitigated, such biases risk producing unfair or discriminatory outcomes, disproportionately affecting marginalized groups, propagating misinformation, and eroding user trust. These risks are further intensified by the large-scale deployment of LLMs, which could accelerate harmful societal feedback loops.

Social biases are commonly understood as leading to representational harms (e.g., misrepresentation, stereotyping, uneven model performance, use of derogatory language, reinforcement of exclusionary norms) and allocative harms (denial of resources, opportunities, or equitable treatment, either directly or indirectly) [115]. Their origins are multifaceted, encompassing skewed or unrepresentative datasets, inconsistent annotation practices, architectural and loss function choices, as well as implicit human biases rooted in the cultural and social backgrounds of developers.

Recent surveys classify bias into five main categories along the NLP pipeline [153, 230]: data, annotation, representation, model, and research design. Data bias emerges from unrepresentative corpora; annotation bias reflects inconsistencies or subjectivities in labeling; representation bias is introduced through the encoding of inputs into embeddings; model bias arises from optimization and architectural choices; and research design bias stems from higher-level methodological decisions, such as benchmark selection and problem framing. This taxonomy complements outcome-oriented classifica-

tions (e.g., demographic or systemic bias) by emphasizing the points of origin where mitigation strategies can be applied.

Bias detection and measurement combine qualitative methods (e.g., surveys, interviews, user studies) with quantitative metrics such as error rate disparities, equal opportunity, statistical parity, or equalized odds [372]. However, LLMs increase the complexity of this task, as bias often exhibits a multi-dimensional nature arising from interactions among multiple features. To address this complexity, composite measures such as the Large Language Model Bias Index (LLMBI) [249] have been introduced, integrating multiple dimensions of bias (e.g., age, gender, and race) into a single score. Moreover, a range of benchmark datasets and evaluation frameworks has been introduced to facilitate the systematic assessment of bias in LLMs [33, 209, 351]. While detection is a prerequisite, mitigation strategies must translate these insights into corrective interventions.

Mitigation approaches are typically applied at different stages of the model lifecycle [115, 230]. Pre-processing strategies target the data (augmentation, filtering, re-weighting, synthetic generation, instruction tuning). During training, model parameters, architectures, and loss functions may be adapted, for example through adversarial or contrastive objectives, selective fine-tuning, or projection-based debiasing. Post-training and inference-time interventions adjust decoding strategies, redistribute token probabilities, or incorporate modular debiasing components [137, 183, 229]. Post-processing further refines outputs, e.g. via keyword replacement or translation-based rewriting [18, 90].

Despite this broad toolkit, trade-offs are unavoidable: bias mitigation often comes at the expense of overall accuracy, scalability, or cross-domain robustness. Many methods target a single bias dimension (e.g. gender), but

perform less effectively in intersectional settings where identities overlap (e.g. gender × race × class). Moreover, mitigation in LLMs is particularly complex, as biases are multi-dimensional and may resurface when models are deployed in new cultural or linguistic contexts.

Beyond text-only NLP, biases also affect multimodal models (vision-language, speech-language), where correlations between modalities (e.g. gendered occupation labels aligned with stereotypical images) introduce new challenges [351]. Bias is also dynamic: it evolves with deployment, as user interactions and feedback loops can reinforce prejudices over time, while shifts across domains or languages expose fairness gaps that remain invisible in controlled benchmarks.

Finally, it is important to acknowledge that fairness is not only a technical concept. Competing philosophical frameworks, such as egalitarian versus utilitarian approaches to fairness [44, 99], and legal requirements, including Equal Opportunity law, the GDPR [346], and the forthcoming EU AI Act [102], shape the very definition of what counts as "fair." In addition to technical mitigation, socio-technical strategies such as dataset governance, external audits [276], participatory design, and stakeholder engagement [34] are increasingly recognized as necessary complements to purely algorithmic solutions.

Transparency and Explanability

Recent advances in artificial intelligence have led to substantial improvements in model performance, albeit often reducing interpretability. Many AI systems operate as black boxes, where the internal mechanisms and decision-making processes is opaque and difficult for human users to analyse. This lack of transparency undermines trust in everyday applications such as

conversational agents, recommendation systems, and decision-support tools. eXplainable AI (XAI) field addresses this challenge by developing methods and interfaces that make model behaviour understandable and trustworthy for humans [80, 190, 290, 345].

We group XAI methods along four axes. (i) Ante-hoc (intrinsic) vs. post-hoc: intrinsic methods enforce interpretability by design (e.g., sparse linear models, rule lists, prototype-based models, concept bottlenecks), whereas post-hoc methods generate explanations after prediction. (ii) Local vs. global: local explanations account for individual predictions, while global explanations characterise the model's overall decision-making logic. (iii) Model-specific vs. model-agnostic: methods that leverage a particular architecture versus those applicable across models. (iv) Derivation vs. presentation: mathematically principled procedures produce raw explanations, which are subsequently communicated via human-centred visual or textual interfaces.

Among post-hoc methods, feature-attribution approaches assign importance scores to input features such as tokens, n-grams, or learned representations. Common examples include gradient-based saliency maps and their refinements (e.g., Integrated Gradients [332], DeepLIFT [323]), as well as attention visualisation [86]. However, attention weights are not explanations per se and must be interpreted with caution.

Model-agnostic surrogates such as LIME [284] and SHAP [216] approximate local or global behaviour by fitting an interpretable proxy around the black box. Specifically, LIME architecture perturbs inputs, collects model predictions, and trains a sparse linear model to approximate local behaviour. SP-LIME extends this by selecting diverse representative instances to offer a global view. Similarly, SHAP provides a post-hoc, model-agnostic

approach grounded in cooperative game theory's Shapley values, ensuring local accuracy, null effects for absent features, and consistency. Variants include Kernel SHAP (model-agnostic sampling) and Deep SHAP (fast approximations for neural networks), unifying approaches such as LIME and DeepLIFT [323] under a single theoretical framework. Despite their popularity, these techniques face challenges concerning stability, faithfulness under feature correlation, and sensitivity to the choice of perturbations.

Complementary to attribution-based methods are example-driven and provenance-based explanations. The former retrieve semantically similar instances or influential training points while the latter reconstruct multi-step reasoning chains or data lineage when available [267]. Declarative approaches further induce human-readable artefacts such as rules, trees, or programs [81, 302]. In contrast, intrinsic or ante-hoc architectures aim to expose interpretable components within the model itself. Examples include Self-Explaining Neural Networks (SENN) [16], concept bottleneck models [187], and prototype-based networks [274] that link predictions to representative prototypes. Such models aim to improve faithfulness by aligning internal computations with semantically meaningful factors.

Within NLP, LLMs increasingly generate natural-language explanations (i.e. rationales) jointly with predictions, either in extractive or abstractive form [380]. Reasoning-oriented prompting strategies (e.g., Chain-of-Thought [299] or ReAct [365]) decompose complex tasks into intermediate steps, thereby facilitating both performance and interpretability. Nevertheless, natural-language rationales are not guaranteed to faithfully represent underlying computations, and their evaluation requires fidelity-aware metrics in addition to plausibility.

The presentation of explanations to end users constitutes a further cru-

cial dimension. Saliency maps, attention plots, counterfactual interfaces, rule or provenance visualisations, and hybrid textual–symbolic formats all provide complementary ways of communicating explanations, with the choice of medium ideally tailored to the expertise of the intended stakeholders [31, 80]. Finally, rigorous evaluation of XAI methods necessitates a clear distinction between faithfulness (i.e. the degree to which an explanation reflects the true behaviour of the model) and plausibility (i.e. the extent to which it is convincing to human observers). Additional criteria include stability, completeness, and usability [160]. Benchmarks such as ERASER [89] facilitate rationale-based evaluation, while sanity checks test the robustness of attribution techniques. In the case of surrogate models, ensuring high fidelity between proxy and original system is of particular importance.

Accountability

In the Ethical AI field, the accountability principle refers to the responsibility for decisions and actions performed by intelligent systems. It is essential for safeguarding fairness, preventing harm, and fostering public trust in emerging technologies [156, 175].

Given its close connection to the principles of fairness and transparency, the development of accountable AI introduces a set of challenges that partly overlap with those previously discussed. Chief among these is the black-box nature of modern machine learning models, whose internal logic remains opaque even to their developers. While earlier sections addressed this opacity primarily in terms of interpretability and user trust, here it is examined through the lens of accountability. In this context, the lack of transparency contributes to what has been termed the "problem of many hands" [156, 171], in which responsibility for outcomes becomes diffused

across the multiple stakeholders involved in the design, deployment, and operation of AI systems. Responsibility is therefore fragmented among different actors (e.g., programmers, data providers, operators, and end-users), which makes the attribution of liability in cases of malfunction or harm particularly complex. This difficulty is compounded by the absence of self-awareness and moral agency in AI systems, which prevents them from being considered responsible agents within ethical or legal frameworks. The resulting diffusion of responsibility often weakens the sense of accountability among human stakeholders, especially when algorithmic outputs are accepted uncritically or when multiple intermediaries are involved. Addressing these challenges requires the explicit definition and assignment of liability throughout the technology's lifecycle. Designers, developers, owners, operators, and end users should all be recognized as responsible stakeholders, both before and after deployment. To make accountability operational, methodological tools such as audit trails and activity logs [319] have been developed. These allow retrospective analyses of errors, near misses, and usage patterns, thereby supporting the identification of responsibility when adverse outcomes occur.

The black-box nature of AI systems also highlights human-limited control over models, leading to the emerging field of controllable AI [177, 205]. Section 2.2 (see Chapter 5) offers an in-depth analysis of this field, with a focus on grammar-based methods as a specific branch of controllable AI.

Another key dimension of accountability concerns privacy and security. AI systems rely on large volumes of data, which entails risks of unauthorized access, breaches, and unethical exploitation [171]. In this regard, the EU's General Data Protection Regulation (GDPR) highlights higher compliance standards for organizations and incentivizes the diffusion of privacy-

enhancing technologies (PETs) to minimize disclosure risks [184]. Particularly relevant are issues of informed consent and data ownership: individuals should retain meaningful control over their personal information and a clear understanding of how it is collected, processed, and used. Yet users often lack awareness of the implications of data sharing, giving rise to the so-called privacy paradox [188], in which strong concerns about privacy coexist with limited protective behaviour. Addressing these challenges requires coordinated action among developers, regulators, and end users. Solutions include robust security policies and the adoption of PETs such as homomorphic encryption, federated learning, and differential privacy (DP). DP in particular provides a formal mathematical framework that ensures that the output of an AI model does not depend on the data of any single individual.

Recent research has advanced DP methodologies through optimization strategies that mitigate bias during training, thereby enhancing both fairness and privacy protection [271]. Other innovations include adaptive friction mechanisms for optimizers [374] and gradient-noise algorithms [212], which improve training efficiency while maintaining protection of sensitive information. Alongside DP, further approaches include federated learning, which enables distributed training without centralizing data [176]; anonymization and pseudonymization techniques that remove identifiable entities while preserving semantics [219]; secure multi-party computation and homomorphic encryption, which allow computations on encrypted data [107]; and adversarial training methods designed to reduce the memorization of sensitive details [125].

1.3 Human-Centered Approaches to Designing and Evaluating AI

In the previous section (cf. Sect. 1.2), we examined the principles of HCAI, outlining their definition and reviewing state-of-the-art approaches in NLP. Building on this foundation, the current section describes the design process for integrating such systems into human-centered ideation workflows. Our focus is on methodologies such as Value-Sensitive Design (VSD), Participatory Design (PD), and comprehensive user evaluations. Collectively, these approaches promote the development of AI systems that not only prioritize human well-being but also remain aligned with real-world contexts and needs [65, 269].

Specifically, Section 1.3.1examines human-oriented design methods, with particular attention to personalization, VSD, and PD that actively involve stakeholders in shaping AI technologies. In Section 1.3.2, we discuss evaluation frameworks for AI systems in the HCAI field.

1.3.1 Human-oriented Design Methods

Shifting the emphasis from efficiency-driven automation to the amplification of human capabilities and experiences, the HCAI-oriented design framework aims to adapt AI systems to human needs, avoiding the imposition of rigid technological constraints on individuals [117, 269, 320, 361]. This paradigm highlights the importance of a deep understanding of the different needs, values, and contexts of users and stakeholders.

Within human-oriented design methods, we focus on two main areas: approaches addressing emotions and values, exemplified by VSD and PD, and approaches focused on inclusivity. Section 1.3.1 discusses the first area,

highlighting methods that integrate emotions and values into the design process; while Section 1.3.1 focuses on inclusivity as a critical dimension of human-oriented design.

Emotion- and Value-aware Approaches

Human-centered AI systems are increasingly designed to recognize and respond to human emotions, with the goal of fostering empathetic interactions, supporting emotional regulation, and strengthening social connectedness. Adaptive user models, derived from the analysis of behaviors and preferences, enable dynamic system adjustments to improve contextual relevance and user experience. This perspective aligns with psychological theories that emphasize autonomy, competence, and relatedness as core human needs.

Within the broader tradition of Human-Centered Design [75], two influential methodological approaches have attained particular prominence in both research and practice. VSD offers systematic methodologies for embedding human values into technology [340], while PD emphasizes the active involvement of stakeholders in design and development processes [329]. Together, they provide complementary strategies to ensure that AI systems reflect ethical, cultural, and individual values, while supporting principles such as fairness, transparency, and human rights.

Empirical research illustrates the benefits of incorporating emotional and value-based dimensions into AI. Emotion-aware recommender systems, for example, can suggest music based on users' real-time affective states, enhancing perceived relevance [30]. Similarly, ERS-IEQ adapts indoor environmental quality according to users' emotions by employing the RECS ontology [181]. Other methods, such as the Affective Index and Affective Index Indicators (AII), analyze text data to estimate emotional probabili-

ties and generate affect-based recommendations while maintaining privacy protection [201].

Industrial applications reflect similar trends. Microsoft Cortana has patented mechanisms for inferring emotional states from contextual cues and prior interactions, aiming to deliver more adaptive responses[4]. Mastercard's Shopping Muse integrates NLP and image recognition to interpret colloquial queries and provide personalized product suggestions[5]. While these developments highlight the commercial relevance of empathy-enabled AI, they also raise concerns regarding robustness, transparency, and user trust, especially when applied in consumer contexts where privacy is critical.

Value-aware personalization extends this perspective. Initiatives such as VALAWAI (Value-Aware Artificial Intelligence), funded by the European Innovation Council, are developing toolkits for value-sensitive AI inspired by Global Neuronal Workspace models and informed by moral decision-making theories[6]. Hybrid human–AI systems pursue similar objectives by combining algorithmic efficiency with human empathy and contextual reasoning, with human agents providing ethical oversight[7].

Applications are emerging across domains. In cultural heritage, the CUPIDO system employs VR and AI to detect visitors' emotions and values, enabling cross-cultural comparison and collective reflection on artworks, thereby fostering inclusion and participation [54]. In education, AI-based tools for social-emotional learning promote empathy, self-management, and interpersonal skills through simulations and adaptive feedback [239]. Platforms such as Coursera experiment with the analysis of facial expressions

[4]https://www.windowscentral.com/cortana-could-simulate-emotional-empathy-according-patent

[5]https://www.mastercard.com/us/en/news-and-trends/Insights/2025/ai-and-personalization-can-close-the-empathy-gap-between-brands-and-their-customers.html

[6]https://cordis.europa.eu/project/id/101070930

[7]https://www.v2solutions.com/whitepapers/hybrid-ai-human-personalization/

and engagement metrics to adapt content delivery, aiming to increase personalization and learning effectiveness.

Overall, embedding emotional, social, and value-oriented dimensions into AI design shifts the focus from purely technical optimization to human well-being and responsible alignment with societal values. At the same time, challenges remain: emotion detection could be affected to bias and cultural variability; personalization may reinforce stereotypes; and the ethical governance of value-sensitive systems requires further empirical validation (cf. Sect. 1.2). Addressing these open issues is essential to ensure that HCAI contributes not only to innovation but also to inclusive and sustainable human flourishing.

Inclusivity-oriented Approaches

Recent advances in NLP highlight the growing importance of inclusivity, expanding the scope of language technologies beyond conventional text and speech in order to address diverse modes of human communication. While state-of-the-art NLP models achieve considerable performance in tasks such as translation, summarization, and dialogue generation, they frequently reflect biases embedded in predominant linguistic and cultural corpora, thereby marginalizing communities whose communicative practices diverge from the mainstream. To address this challenge, an emerging strand of computational linguistics conceptualizes machine learning systems as instruments for social good, explicitly oriented toward enhancing accessibility.

A central research area in HCAI for accessibility is the development of sign language technologies. Computational methods for sign language recognition, translation, and representation have advanced considerably in recent years. Surveys of state-of-the-art approaches (covering recognition, trans-

lation, and avatar-based representation models) highlight both substantial progress and ongoing challenges, particularly regarding dataset scarcity, sensor limitations, and the inherently multimodal nature of sign languages [255]. Chapter 6 provides an in-depth evaluation of a specific task within this field, while Section 2.3 introduces the state-of-the-art method for sign language processing based on a dedicated encoding language.

Parallel research efforts support accessibility for individuals with visual impairments. Automatic image captioning systems have been applied to generate textual descriptions of visual content, facilitating perception for people who are blind or have low vision. Techniques leveraging pretrained CNNs and LSTM-based RNNs can automatically generate descriptive captions from raw images, enhancing usability in educational and digital environments [71]. Systematic reviews of touchscreen-based interfaces for screen reader users highlight both promising prototypes and critical gaps (e.g. lack of automated content generation and limited involvement of end users in design) underscoring the need for more user-centered, scalable approaches [248].

These research directions show how advances in NLP support the development of inclusive AI technologies, enabling richer linguistic representations and more robust models.

1.3.2 Evaluation of AI

Evaluation is a central stage in the development of HCAI models. Traditional AI assessment has largely focused on quantitative indicators such as accuracy, precision, or computational efficiency. While essential, these metrics alone are insufficient to capture the complex interactions between AI systems and human users. From a human-oriented perspective, evaluation

must also consider trust, usability, fairness, transparency, and broader societal implications. This view recognizes that AI systems are not isolated technical artifacts but socio-technical constructs whose value depends on their alignment with human needs, values, and contexts [17, 320].

In this framework, evaluation is no longer a terminal step after deployment. Instead, it becomes an iterative and participatory process integrated throughout the AI lifecycle. Continuous evaluation allows developers to identify risks early, anticipate unintended consequences, and adapt systems in ways that promote user empowerment rather than technological determinism. This approach connects engineering-driven metrics with human-centred methodologies, laying the foundation for responsible and sustainable AI practices.

A human-centred perspective requires involving participants in assessing subjective qualities such as comfort, trust, and perceived fairness [117, 278]. Such participation reduces the need for costly late-stage corrections. Garibay et al. [117] propose a framework structured around four dimensions: people, process, product, and principles. It emphasizes grounding evaluation in user needs and values (people), adopting observation, usability testing, and iterative refinement (process), and ensuring that design methods enhance human capabilities while preserving user control (principles). Importantly, this framework broadens the notion of usability to include issues such as undue influence, automation errors, or dynamic reliability. The ultimate aim is to design systems that are understandable, explainable, and support deliberate trust calibration.

Explainability and Fairness provide clear examples of this multidimensional evaluation. Assessing whether a system is explainable, validating both XAI models (cf. Sect.1.2.3), and conforms to the fairness principle (cf.

Sect. 1.2.3) requires both qualitative and quantitative methods [42, 132, 290]. In the XAI field, qualitative approaches focus on human-centric aspects such as comprehensibility, user understanding, and usability. These are often assessed through user studies using proxy tasks like forward simulation or counterfactual reasoning. In contrast, quantitative approaches measure dimensions such as faithfulness (the match between explanations and internal model reasoning), plausibility (the perceived reasonableness of explanations), and the linguistic quality of natural language explanations (NLEs). Benchmark datasets such as ERASER [89], e-SNLI [64], and HateXplain [226] provide human-annotated rationales that support systematic validation.

In the bias detection and Fairness model's evaluation fields, evaluation relies on three main categories of metrics: embedding-based, probability-based, and text-based [115]. These metrics capture vector space distances between concepts, token likelihoods when protected attributes are perturbed, or undesirable patterns such as toxicity in generated text. Common evaluation strategies include counterfactual input datasets (pairs of sentences with altered social groups) and generative prompts designed to elicit biased outputs. Specific benchmarks such as SBIC [300], and Winobias [373], often combined with similarity metrics like ROUGE or BERTScore, are widely used to assess alignment with human moral judgments.

Evaluation also concerns the system's capacity to align with moral values and social norms, including robustness against adversarial inputs, ambiguous scenarios, or toxic content. Reliability and privacy protections become essential to ensuring safe and responsible use. We examine these aspects in Chapter 2, where we discuss benchmark datasets, evaluation theories, and models that have achieved state-of-the-art performance for the moral values

detection tasks.

Finally, HCAI evaluation methodologies increasingly rely on user studies, interviews, focus groups, and simulations to collect evidence about user preferences, comprehension, and contextual factors. These data help refine explanations and behaviours, ensuring that both XAI solutions and AI systems more broadly address the concrete and inclusive needs of their users.

1.4 Humans Teaming with AI

Human–AI teaming refers to systems in which humans and AI agents collaborate in order to achieve outcomes that neither could accomplish independently. Evaluations of these systems commonly focus on two key dimensions: objective performance improvements, measured at the level of the human, the AI, or the joint team; and the subjective quality of the collaboration as experienced by the human partner. Conceptually, these approaches can be situated along a spectrum of human–AI integration. At one extreme, AI systems operate as surrogates for humans, thereby reducing or even displacing human agency. In intermediate configurations, humans remain in the loop, supervising, guiding, or correcting automated processes. At the opposite extreme, humans and AI function as teammates, engaging in reciprocal interaction and collaborative decision-making.

The most collaborative paradigm is often characterized as human–machine teaming (HMT) [218]. In this model, humans and AI function as an integrated unit, drawing upon principles of human–human teamwork, including shared goals, mutual situational awareness, trust, and bidirectional communication. From a design perspective, the AI is expected to act as a competent teammate rather than as a mere tool [65, 361]. Concrete examples include language-related domains where human interpretative skills complement au-

tomated capabilities (fact-checking, journalism, unstructured text analysis, or cross-language code translation). In such contexts, the system efficency do not related in its computational capacity but is focused on its ability to enable fluid, cooperative workflows.

A related paradigm is humans in the loop (HITL), which emphasizes sustained human involvement in the development, training, and refinement of AI systems [21, 38, 39, 43, 82, 100, 141, 143, 144]. In this framework, common tasks include dataset annotation, error correction, and iterative model evaluation. Importantly, user interaction with model outputs can itself generate new inputs that shape system evolution. HITL practices have gained importance with the emergence of LLMs, whose opacity and potential unreliability make sustained human oversight essential. This requirement aligns closely with the broader paradigm of HCAI, which emphasizes human control, accountability, and ethical responsibility as foundational principles of system design.

Beyond supervised model development, human–AI collaboration is also prominent in the domain of automated machine learning (AutoML) [240]. AutoML frameworks aim to streamline the ML pipeline by automating tasks such as data preprocessing, feature engineering, and model selection [348, 358]. Rather than displacing human expertise, these systems seek to augment it by embedding automation within pipelines where human judgment remains decisive. Their claim to human-centeredness cames from this synergy: automation is valuable only insofar as it extends the scope of expert decision-making rather than constraining it.

At the opposite extreme are AI systems conceived as human substitutes, intended to replicate or replace human performance in specific domains such as diagnostic imaging or automated grading. These approaches often neglect

principles of HCAI, sidelining human interpretative capacity and agency.

Collectively, these paradigms present a dicothomy: whether AI should replicate, assist, or collaborate with humans. Current research increasingly indicates that the most durable advances arise from collaborative models that preserve human oversight, judgment, and responsibility.

2 Human-centered AI Methods: State of the art

This chapter presents a comprehensive review of the state of the art in HCAI, with particular emphasis on the role of moral values and the development of assessable, controllable, and inclusive models.

The first part delves into the literature related to the classification of moral values according to well-defined theoretical frameworks such as Haidt's Moral Foundations Theory [129], Schwartz's Theory of Basic Human Values [313], or Curry's Morality-as-Cooperation framework [78]. For each of these frameworks, we focus on the main benchmark datasets and lexicons developed for value-based text annotation, together with the configured machine learning methods (cf. Sect. 3).

Then, we discuss two challenges involved in building assessable and controllable models for a trustworthy AI, focusing on the issue on evaluation metrics adapted for high degree of subjectivity tasks, as emotion or moral values detection (cf. Sect. 2.2.1). In Section 2.2.2, we present the state-of-the art for grammar-constrained methodologies for generative models, a specific group of methods that deal with the challenge of output controllability in LLMs (cf. Sect. 2.2.2) in order to have output more aligned with human requirements.

Finally, we focus on an emerging application area of great relevance for

HCAI: AI for Inclusivity (cf. Sect. 2.3). In this section, we introduce the topic of automatic bidirectional translation between natural language and sign language with a specific focus on the main sign language encodings used for these methods.

2.1 Moral Values Detection

Automatic moral classification in text is a rapidly evolving field with significant implications for NLP, social sciences, and ethical decision-making. The current state of the art has increasingly relied on well-established theoretical frameworks (i.e. Haidt's Moral Foundations Theory [129], Schwartz's Theory of Basic Human Values [313], and Curry's Morality-as-Cooperation[78]) to guide the development of benchmark datasets, annotation schemes, and computational lexicons. These resources support the development and evaluation of both supervised and unsupervised approaches for moral value classification, ranging from fine-tuned Transformer architectures to prompting-based strategies with LLMs.

To provide an overview of the field, Section 2.1.1 reviews benchmark datasets and state-of-the-art models developed within Haidt's Moral Foundations Theory [129]. Section 2.1.2 highlights key contributions grounded in Schwartz's Theory of Basic Human Values [313], while Section 2.1.3 discusses recent datasets and models based on Curry's Morality-as-Cooperation framework [78].

2.1.1 Moral Foundation Theory

According to MFT, moral judgments vary significantly throughout cultures and over time, but there are core moral dimensions that form the foundation of an "intuitive" ethical system across human societies. This theory

integrates four key perspectives on the origins and nature of morality: nativism, cultural development, intuitionism, and pluralism. MFT acknowledges the potential role of neurophysiological bases for moral responses (nativism), recognizes the influence of environmental factors on moral development (cultural-evolutionary), suggests that moral judgments arise from various cognitive processes (intuitionism), and allows for the existence of multiple moral frameworks across cultures (pluralism) [129]. At the core of MFT there are five distinct moral dimensions or dyads, each representing a fundamental dimension of human moral reasoning (Figure 2.1):

- **Care/Harm**: This domain highlights the importance of avoiding harm and promoting the well-being of others. It is associated with virtues such as kindness, nurturance, and compassion.

- **Fairness/Cheating**: This domain focuses on the concepts of justice, equity, and reciprocity. It supports the notion of fairness in social interactions and promotes cooperation.

- **Loyalty/Betrayal**: This domain highlights the importance of loyalty to one's group and disapproval of betrayal. It promotes social cohesion and cooperation within groups.

- **Authority/Subversion**: This domain promotes respect for legitimate authority, social order, and hierarchies. It promotes adherence to rules and social norms.

- **Purity/Degradation**: This domain focuses on ideas of cleanliness and contamination, often extending to moral concepts. It can be linked to notions as spiritual purity and adherence to social taboos.

- **Liberty/Oppression**: This dimension, captures individuals' sensi-

Figure 2.1: MFT framework

tivity to issues of personal freedom and resistance to domination or coercion by others. It is associated with values such as autonomy, individual rights, and opposition to tyranny.

The five core moral dimensions constitute the foundation of MFT, offering a comprehensive framework for analyzing the multifaceted nature of human morality. Subsequent refinements to MFT introduced the "Liberty/Oppression" dimension and differentiated the "Fairness" dyad into "Equality" and "Proportionality" [27].

Section 2.1.1 introduces the benchmark datasets developed within this theoretical framework, which serve as the basis for evaluating AI-based moral value classifiers. In Section 2.1.1, we present state-of-the-art models across two methodological paradigms (i.e. in-distribution and out-of-distribution) scenarios. We further detail current methodologies that using LLMs as zero-shot unsupervised moral values classifiers through different prompting strategies.

Dataset

Research on MFT has produced a diverse ecosystem of resources, ranging from annotated corpora to psychometric surveys and lexical tools, each enabling the study of moral values from complementary perspectives and modalities [369]: annotated corpora capture moral language in naturally occurring discourse; psychometric instruments provide standardized benchmarks of individual moral orientation; and lexical resources support scalable computational analysis of moral categories (Table 2.1).

Among the most widely used corpora, many adopt a dyadic labeling structure, treating each foundation as a pair of opposing values (e.g., Care vs. Harm), thereby enabling polarity-sensitive analysis. The Moral Foundations Twitter Corpus (MFTC) [148] reflects this approach, comprising tweets labeled across ten moral categories plus a non-moral class, spanning seven topical domains such as All Lives Matter/Black Lives Matter and the #MeToo movement. Similarly, Beiro et al. [36], Rojecki et al. [289] and, Pacheco et al.[254] present a corpus of 4,498, 2,648, and 750, respectively, on the COVID-19 vaccination.

Political discourse has also been labeled within this framework. Johnson et al. [169] present a dataset of 2,050 tweets from U.S. politicians across six domains, annotated with ten different moral foundation values for topic and policy frame. Roy et al. [291] expand this corpus through lexicon-based retrieval, collecting a total of 161,295 items. In the Italian context, MoralConvITA [330] focuses on immigration-related Twitter conversations, preserving the dyadic moral categories while introducing stance labels (support, attack, continuation) to capture interactional dynamics within tweet–reply pairs.

Among corpora that extend or adapt the standard MFT schema, Trager et al. introduced the Moral Foundations Reddit Corpus (MFRC) [338],

which contains 16,123 Reddit comments annotated with a revised set of moral categories. This framework refines the conventional MFT taxonomy in two key ways. First, it decomposes the Fairness/Cheating foundation into two distinct subdimensions: Equality/Inequality and Proportionality/Disproportionality [25]. Second, it introduces a novel category, Thin Morality, to capture moral judgments that are either too vague or context-dependent to be unequivocally assigned to a standard foundation. Finally, all content devoid of moral reasoning is explicitly categorized under a Non-moral label. The Moral Foundations News Corpus (MNFC) [352] applies moral annotation to news articles, while the Moral Integrity Corpus (MIC) [378] enriches the Social Chemistry 101 dataset of social norms [113] with moral labels for everyday situations. In the argumentative domain, the Extended ArgQuality Corpus [185] links moral values to argument quality and stance, whereas MoralArg [13] provides a large-scale argumentative dataset explicitly balanced across foundations. Beyond text, narrative and multimodal resources such as the Moral and Affective Film Set (MAAFS) [228] and the Moral Foundations Vignettes (MFVs) [74] extend moral content analysis to audiovisual and controlled textual stimuli.

More recent contributions emphasize multilinguality and interpretability. MFTCXplain [343] introduces a corpus in English, Italian, Persian, and Portuguese, focused on hate speech contexts. Each tweet is annotated with up to three moral categories and enriched with span-level rationales that support multi-hop explanations, facilitating research on explainable moral classification. The Event-level Moral Opinions in News Articles (EMONA) dataset [199] shifts attention from social media to journalism, providing fine-grained event-level moral annotation. EMONA comprises 400 news articles (45,199 event mentions, 9,613 annotated with moral judgments) sourced

from AllSides (180), BASIL (150), and MPQA 3.0 (70). Annotation proceeds in two stages: first identifying event mentions, then assigning moral or "non-moral" labels using both local sentential and broader document context. This design enables the capture of implicit moral judgments, which are often subtle in news reporting, and expands the scope of moral annotation beyond short, user-generated texts.

Complementing annotated corpora, which infer moral sentiment from text, psychometric instruments provide direct measurements of an individual's moral orientations. The widely used Moral Foundations Questionnaire (MFQ) [128] operationalizes the original theory through 30 items on a 6-point Likert scale [206]. Its successor, MFQ-2 [25], introduces a critical refinement by disambiguating the broad Fairness foundation into two distinct subscales: Equality and Proportionality dimensions. This expansion from five to six foundations allows for finer-grained, more cross-culturally valid assessments of fairness perceptions.

In parallel, lexical resources facilitate large-scale computational analysis of moral language. The Moral Foundations Dictionary (MFD) [126] contains 324 stems and lemmas (32 per category, on average), later expanded to 2,014 entries in MFDv2 [114]. The extended MFD (eMFD) [150], derived from MNFC, introduces probabilistic weighting based on contextual usage. MoralStrength [22] expands coverage with 1,000 lemmas and adds graded scores for both moral relevance and intensity, while LibertyMFD [23] incorporates an additional Liberty category. Domain-specific lexicons, such as those developed by Roy et al. [291] for gun control and immigration, further demonstrate the adaptability of moral lexicons to targeted policy areas.

Table 2.1: Overview of benchmark moral surveys, lexicons, and datasets for moral value detection within the MFT framework. For each resource, we report its category, size, number of labels, the employed taxonomy (i.e., moral foundations values (MFVs) or dimensions (MFDs)), extensions beyond the original MFT (e.g., Liberty/Oppression (L-O), Equality/Proportionality (E, P), Thin Morality (T), Non-Moral (NM), or Hate Speech (H)), and the number of annotators.

Moral Survey

Dataset	Size	Label	Taxonomy	Extra Label	Annotation
MFQ [128]	30	5	MFDs	NM	–
MFQ-2 [26]	36	6	MFDs	E, P	–

Moral Lexicons

Dataset	Size	Label	Taxonomy	Extra Label	Annotation
MFDv1 [127]	324	10	MFVs	–	–
MFDv2	2,014	10	MFVs	–	–
eMFD [151]	3,270	5	MFDs	–	–
MoralStrength [22]	1,000	5	MFDs	–	$\geq$5

MFT-Annotated Datasets

Dataset	Size	Label	Taxonomy	Extra Label	Annotation
MFTC [148]	35,108	10	MFVs	NM	$\geq$3
MFRC [338]	16,123	7	MFDs	E, P, T, NM	$\geq$3
MFNC [352]	35,985	10	MFVs	–	557
Covid [289]	2,648	10	MFVs	–	–
Congress [169]	2,050	10	MFVs	–	2
Ext. Congress [291]	161,295	10	MFVs	–	–
Facebook [36]	4,498	12	MFVs	L-O	9
Covid-19 [254]	750	5	MFDs	–	3
Ext. ArgQuality [185]	320	5	MFDs	–	2
MoralArg [13]	230,000	5	MFDs	–	–
MIC [378]	38,000	5	MFDs	–	3
MAAFS [228]	69	6	MFDs	L	575
MFVs [74]	132	6	MFDs	L	510
MFTCXplain [343]	3,000	6	MFDs	H	5
EMONA [199]	45,199	10	MFVs	NM	5

Methods and Results

We group moral foundation classification models into two categories: moral-driven and moral-targeted pretrained language models (PLMs) [369]. Moral-driven approaches rely on explicit supervision with moral foundation annotations, while moral-targeted approaches probe implicit moral knowledge encoded during pretraining, typically through prompting without task-specific fine-tuning.

Among the moral-driven approaches, supervised fine-tuning techniques remain the most established paradigm. Models evaluation is typically conducted on standard benchmarks (cf. Table 2.1) across three different settings: in-domain (ID), where models are trained and tested on data from the same distribution; out-of-domain (OOD), where the test data originates from a different distribution than the training data, though the moral taxonomy is consistent; and cross-domain, where evaluations are performed on distinct topical subcorpora within the same dataset.

In contrast, moral-targeted approaches operate without explicit MFT supervision. Instead, they directly probe the intrinsic moral knowledge encoded within PLMs, without any task-specific fine-tuning. This paradigm probes the moral reasoning acquired during pretraining by analyzing model generations for morally charged prompts. Consequently, MFT serves not as a supervisory signal but as an ex post analytical lens for interpreting model responses. This research direction utilizes zero-shot prompting (i.e. in-context learning) with both LLMs and smaller PLMs. These methods often demonstrate superior generalization by avoiding overfitting to narrow training distributions; instead, they leverage the commonsense and culturally nuanced moral knowledge acquired during pretraining [56].

A complementary line of research investigates moral biases encoded within

contextual embeddings, revealing that their geometric spaces often capture culturally contingent or systematically skewed moral assumptions.

Moral-driven Approaches: Methods Within the moral-driven category, transformer-based models are frequently adapted for moral classification. Trager et al. [338], Hoover et al. [148], and Ziems et al. [378] fine-tune BERT-based models for multi-label classification, while Nguyen et al. [243] train binary classifiers for each of the five moral foundations. Lei et al. [199] extend Longformer with a Bi-LSTM layer to better capture article-level context. Lexicon-based methods remain relevant, such as WN-PPR [185], which enriches the Moral Foundations Dictionary through WordNet-based sense disambiguation and Personalized PageRank, and SBERT-Wiki, which adapts Sentence-BERT to weakly annotated Wikipedia abstracts.

Beyond standard fine-tuning, several studies explore structured architectures that incorporate additional linguistic, relational, or logical constraints. Johnson et al. [168] approach moral foundation classification in political tweets using Probabilistic Soft Logic (PSL), combining lexical cues, ideological affiliation, political slogans, and policy frames into a hinge-loss Markov random field. This framework captures global dependencies across issues, parties, and discourse strategies, moving beyond unigram-based baselines toward higher-level thematic and ideological reasoning. In parallel, Roy et al. [292] propose Declarative Relation-based Inductive Learning (DRAIL), which, unlike PSL, enables neural components to learn rule weights and supports both local and global inference. In DRAIL, moral foundations are modeled as frame predicates, while moral roles encode entity polarity (e.g., target of care/harm, harm-inducing agent). This reframing shifts the task from text-level classification to joint inference over foundations and roles,

providing a finer-grained account of moral framing in political discourse. Comparing to the Johnson et al. model, key distinction lies in polarity modeling: Johnson and Goldwasser directly classify tweets into positive/negative poles for each foundation, whereas Roy et al. model polarity through entity-level roles, yielding more nuanced analyses of moral evaluations in political communication.

Preniqui et al. [265] introduce MoralBERT, a BERT-based model fine-tuned on a combined corpus of MFTC and MFRC. They present two variants: the standard MoralBERT and an adversarial version (Adv-MoralBERT) designed to learn domain-invariant representations, trained with a weighted loss function to mitigate class imbalance.

Guo et al. [131] present the Domain-Adaptive Moral Foundations (DAMF) architecture, a multi-component framework designed to address the heterogeneity of moral foundation datasets. At its core, a BERT encoder generates contextualized text embeddings, which are further refined through a linear transformation layer to encourage domain invariance. The framework integrates three modules: a moral foundation classifier with a weighted loss function to mitigate class imbalance, an adversarial domain classifier connected via a gradient reversal layer to promote alignment across datasets, and a reconstruction module that preserves semantic integrity against excessive adversarial perturbations. Collectively, these components enable DAMF to learn domain-invariant embeddings that generalize effectively across heterogeneous textual sources, thereby enhancing performance in multi-dataset morality inference.

Zangari et al. introduce ME2-BERT [368], a model that integrates event-based and emotion-aware encodings within a domain-adaptation framework. The architecture relies on an emotion-aware denoising autoencoder com-

bined with contrastive loss and adversarial classification in order to align event-rich and event-poor texts.

Alshomary et al. [13] propose a system for moral argument generation integrated within IBM's Project Debater. Their approach employs a BERT-based classifier, trained using distant supervision on argumentative text, to identify moral foundations. For a given controversial topic and stance, the system retrieves and filters sentences based on their moral framing, subsequently generating coherent arguments that are aligned with the specified moral foundations.

Among hybrid architectures, Sahil et al. [295] introduce a dual-path model that combines RoBERTa embeddings with a graph attention network (GAT) over an extended MFD (GAT-eMFD), fusing these features via cross-attention. They further extend this framework into MOTIV, a multimodal system that incorporates textual, spatial, temporal, and behavioral cues.

Moral-driven Approaches: Results Table 2.2, 2.3 and 2.4 show the overall performance of moral classifiers described above for the benchmark datasets in term of weighted F1 score. Specifically, we report the state-of-the-art performance for the ID, OOD and In-context learning (i.e. zero-shot prompting) settings, showing a comparisons for all moral dimensions and values, when it is possibile. Bold values denote the highest performance among ID models, which are fine-tuned and evaluated on the same data distribution. While these models typically achieve higher scores, their performance is less indicative of generalization. A second bold value highlights the best overall result across OOD and in-context settings. This provides a more reliable measure of generalization, since OOD models are fine-tuned on different distributions and in-context models are evaluated without fine-

tuning. Underlined values indicate the best performance within each individual evaluation setting.

In an ID evaluation setting, the Transformer-based models Trager et al. [338], Hoover et al. [148], and Ziems et al. [378] consistently achieve strong results: Ziems et al. [379] report an overall F1 of 0.76 on MIC (cf. Table 2.2), and Lei et al. [199] reach 0.39 on EMONA (cf. Table 2.4). Notably, GPT-4, even without fine-tuning, attains 0.31 on EMONA, suggesting competitive zero-shot potential.

Even evaluated in a ID setting, the Sahil et al. [295] model exhibits a mixed performance: on the MFTC dataset, their model matches the ID performance reported by Trager et al. [338] and even surpasses it on more challenging dyads like Loyalty (0.68 vs. 0.52) and Purity (0.58 vs. 0.48). However, on the MFRC dataset, their framework fails to outperform Trager et al.'s BERT baseline, which remains the state-of-the-art for ID evaluation across all moral dyads (cf. Table 2.2). Finally, the authors demonstrate their model's applicability on the MOTIV dataset of 1,483 geotagged tweets, where the multimodal design leverages non-textual cues to contextualize moral expression.

Among structured models, Global DRAIL [292] achieves comparable performance with an overall weighted F1 of 0.74, w.r.t. the Johnson et al. [168] PSL-based approach (0.73) on the Congress dataset [168] in a ID setting.

Preniqui et al. [265] evaluate their models in both ID and OOD settings, comparing them with different baselines, as MoralStrength, Word2Vec embeddings with Random Forest, and GPT-4 in a zero-shot setting. In the ID scenario, training is conducted on a merged corpus (MFTC and MFRC), with evaluation splits applied post-merging, making direct comparison to prior work on individual datasets not feaseble (so, we do not report the ID

setting evaluation results in Tables). In this framework, Adv-MoralBERT achieves a macro F1 score of 0.73, outperforming both its standard counterpart and baselines. In an OOD setting, Preniqi et al. test MoralBERT models on the Facebook dataset [36] (cf. Table 2.3) for each MF values, achieving lower performance w.r.t. baseline model, represented to GPT-4, evaluated with an in-context learning strategy (i.e. without any moral-oriented finetuning). Specifically, GPT-4 overperforms the most performing MoralBERT model (i.e. MoralBERT-adv) with an F1-macro of 0.57 w.r.t. the 0.55 achieved by MoralBERT. These results confirm that supervised adaptation techniques achieve superior performance when training and test data distributions are closely aligned; however, their effectiveness diminishes under broader domain shifts, where LLMs benefit from the more generalizable knowledge acquired during large-scale pretraining, in an in-context learning strategy.

In a OOD evaluation setting, Guo et al. [131] evaluate the DAMF generalization, employing different models trained on combination of datasets and testing them in a OOD setting on two different datasets, Congress and Covid, achieving state-of-the-art performance. On the Congress dataset (cf. Table 2.3), DAMF achieves F1-scores of 0.40 when trained on a combination of the Covid and MFTC datasets (i.e. DAMF-v1), surpassing competitor models (i.e. DAMF-v2 and MFTC-DAMF, trained on a combination of the Covid, eMFD, and MFTC datasets and the MFTC, respectively) on 2 point in percentage. On the Covid dataset (cf. Table 2.4), it attains an F1-score of 0.61 with two dataset configurations: Congress and MFTC and Congress, MFTC and MFNC. Finally, authors fine-tune also a BERT-based model on adversarially filtered data to mitigate spurious correlations (i.e. BERT+AFLite), achieving comparable performance on the Covid dataset

when the model is trained on MFTC or on a combination of Congress and MFTC datasets.

Always in a OOD setting, ME2-BERT [368] model achieves state-of-the-art performance on MFTC, MFRC and MFNC corpus, surpassing results achieved by several baselines, including BERT variants fine-tuned on morality-related corpora (E2MoCase, MFTC, MFRC, and their combination), domain-adaptation frameworks such as DAMF [131], and MoralBERT [265], which also adopts domain adaptation. The evaluation further considers lexicon-based systems, such as MoralStrength and DDR, as well as recent LLMs, including Llama-3.1, Gemma-2, and Mistral-Nemo, tested in zero-shot settings. As shown in Table 2.2, ME2-BERT achieves an overall F1 scores of 0.57, 0.51, and 0.26 on MFTC, MFRC, and MFNC, respectively. Performance varies across moral dyads, for MFTC, results range from 0.69 on Care to 0.45 on Purity, while for MFRC they vary between 0.64 and 0.35 for the same categories. On MFNC, overall performance drops substantially, with ME2-BERT reaching an average F1 of 0.26, compared to 0.24 for DAMF, 0.20 for MoralBERT, and 0.21 for DDR. Despite a general performance degradation, ME2-BERT achieves the strongest OOD results. Interestingly, DAMF outperforms ME2-BERT and MoralBERT on Authority, scoring 0.26 against 0.23 and 0.24, respectively. For Fairness, results converge around 0.23 across all three models, while on Loyalty MoralBERT performs comparably to ME2-BERT (both at 0.22). In discriminiating non-moral content, ME2-BERT presents superior performance on the MFTC and MFRC datasets with an F1 score of 0.55 and 0.67, respectively. However, it underperforms on MFNC, where the DDR model achieves the highest F1-score (0.48), surpassing DAMF (0.31) and ME2-BERT (0.37). This suggests that while ME2-BERT generalizes effectively across most moral dimensions,

lexicon-based methods, as the DDR approach, maintain a competitive advantage in scenarios with scarce or implicitly expressed moral content, as exemplified by the MFNC dataset.

When compared with LLMs in a in-context learning strategy, ME2-BERT generally demonstrates superior performance, particularly on MFTC where in-context inference with Gemma-2, Mistral-Nemo, and Llama-3.1 yields overall F1 scores between 0.47 and 0.49 against ME2-BERT's 0.57. On MFRC, GPT-4 achieves the strongest overall performance with an F1 of 0.55, surpassing ME2-BERT (0.51) and excelling specifically in the Care, Authority, and Purity foundations, as well as in moral vs. non-moral classification. Conversely, on MFNC, Mistral-Nemo matches ME2-BERT's overall score (0.26), with LLMs generally demonstrating superior performance across individual moral dyads.

When results are aggregated across datasets, Mistral-Nemo emerges as the most competitive open-source LLM, achieving average F1 scores of 0.49 on MFTC, 0.37 on MFRC, and 0.26 on MFNC. It consistently outperforms counterparts like Llama-3.1 and Gemma-2. Notably, on the challenging MFNC corpus, LLMs surpass ME2-BERT in three of five moral foundations, indicating a degree of robustness to severe domain shift. On the MFRC benchmark GPT-based models demonstrate superior performance. GPT-3.5 achieves an overall F1 of 0.48, while GPT-4 reaches 0.55, surpassing ME2-BERT (0.51). This advantage is particularly pronounced on several dyads, including Care (0.69 vs. 0.64), Fairness (0.53 vs. 0.35), and moral versus non-moral discrimination (0.81 vs. 0.67). Despite these strong results from closed-source models, open LLMs like Mistral-Nemo and Gemma-2 still exhibit a significant performance gap, suffering an average drop of approximately 20% compared to ME2-BERT on MFTC and MFRC. This

underscores the continued competitiveness of specialized, domain-adapted architectures like ME2-BERT. A finer-grained analysis of polarity prediction reveals a complementary trend: specific model strengths vary by moral dimension. For instance, Mistral-Nemo excels at distinguishing Harm from non-moral content, whereas ME2-BERT retains superior effectiveness on categories such as Betrayal, Purity, and Degradation.

Finally, Alshomary et al. [13] evaluate their model on the Extended ArgQuality [185] dataset (cf. Table 2.2) in a OOD setting. Model validation involves an assessment of relevance, coherence, and argumentativeness, supplemented by a user study where liberals and conservatives ranked arguments framed with individualizing, binding, or no moral foundations. For comparison, the authors employ two baselines: a lexicon-based method using MoralStrength and a multi-label BERT model trained on the MFTC, the latter representing the state-of-the-art baseline for the Extended ArgQuality [185] dataset (cf. Table 2.2). Their model outperforms both baselines overall, achieving superior F1 scores in Purity (0.40 vs. 0.28), Authority (0.46 vs. 0.27), and Loyalty (0.34 vs. 0.23). However, for Care and Fairness, the lexicon-based and BERT baselines remain stronger, with the lexicon method showing more consistent performance than the BERT model.

Cross-domain Methods In cross-domain evaluation, state-of-the-art approaches are primarily benchmarked on the MFTC and MFRC datasets, as demonstrated in the studies by Huang et al. [158] and Bulla et al. [55]. Huang et al. [158] introduce the Learning to Adapt Framework (L2AF), a neural architecture designed to mitigate domain shift through instance weighting. The framework combines a Neural Feature Extractor (bi-GRU RNN or BERT) for encoding textual inputs, a Prediction Network for moral

Table 2.2: Overall performance of moral value classifiers on the MFT benchmark datasets. For each dataset, the table reports the F1 scores for the five MFDs, the non-moral category, and the overall score. The training setting of each model is also indicated: ID, OOD, or in-context (i.e. zero-shot). Bold values indicate the highest ID performance and the strongest overall generalization across OOD and in-context settings, whereas underlined values indicate the best performance within each evaluation condition.

MFTC [148]

Models	Care	Fairness	Authority	Loyalty	Sanctity	NM	Total	Setting
BERT [338]	**0.75**	**0.82**	**0.65**	0.52	0.48	-	**0.78**	ID
GAT-RoBERTa [295]	0.62	0.74	0.63	**0.68**	**0.58**	**0.88**	**0.78**	ID
ME2-BERT [368]	**0.69**	**0.67**	0.52	**0.55**	**0.45**	**0.55**	**0.57**	OOD
Gemma-2 [368]	0.64	0.60	0.52	0.47	0.41	0.27	0.49	In-context
Mistral-Nemo [368]	0.63	0.51	0.46	0.48	0.42	0.41	0.49	In-context
Llama-3.1 [368]	0.63	0.61	**0.54**	0.47	**0.45**	0.14	0.47	In-context

MFRC [338]

Models	Care	Fairness	Authority	Loyalty	Sanctity	NM	Total	Setting
GAT-RoBERTa [295]	0.41	0.39	0.25	0.26	0.30	**0.79**	**0.70**	ID
BERT [338]	**0.62**	**0.50**	**0.40**	**0.45**	**0.51**	0.76	-	ID
BERT-E2MoCase [368]	0.48	0.50	0.40	0.35	0.31	0.61	0.44	OOD
ME2-BERT [368]	0.64	**0.59**	**0.49**	0.35	0.36	0.67	0.51	OOD
Mistral-Nemo [368]	0.50	0.51	0.40	0.36	0.23	0.25	0.37	In-context
GPT3.5-double [57]	0.60	0.47	0.27	0.33	**0.73**	0.28	0.48	In-context
GPT4-double [57]	**0.69**	0.53	0.30	**0.38**	0.37	**0.81**	**0.55**	In-context

MFNC [352]

Models	Care	Fairness	Authority	Loyalty	Sanctity	NM	Total	Setting
ME2-BERT [368]	0.31	0.23	0.23	**0.22**	**0.22**	0.37	**0.26**	OOD
DAMF [368]	0.28	0.23	0.26	0.19	0.19	0.31	0.24	OOD
MoralBERT [368]	0.28	0.23	0.24	**0.22**	0.11	0	0.20	OOD
DDR [368]	0.21	0.12	0.19	0.15	0.08	**0.48**	0.21	OOD
LlaMa-3.1 [368]	**0.33**	0.29	0.27	0.14	0.20	0.21	0.24	In-context
Mistral-Nemo [368]	0.31	0.27	0.27	0.17	0.17	0.37	**0.26**	In-context
Gemma-2 [368]	**0.33**	**0.30**	**0.28**	0.21	0.18	0.21	0.25	In-context

MIC [379]

Models	Care	Fairness	Authority	Loyalty	Sanctity	NM	Total	Setting
BERT [379]	0.73	0.56	0.52	0.59	0.37	-	0.76	ID
ALBERT [379]	**0.75**	**0.59**	**0.54**	**0.62**	**0.40**	-	**0.76**	ID

Extended ArgQuality [185]

Models	Care	Fairness	Authority	Loyalty	Sanctity	NM	Total	Setting
Lexicon [14]	**0.60**	0.13	0.23	0.17	0.27	-	0.28	OOD
mBERT [185]	0.50	**0.40**	0.14	0.16	0.21	-	0.28	OOD
WN-PPR + SBERT-Wiki [185]	0.52	0.37	**0.34**	**0.28**	**0.46**	-	**0.40**	OOD

Table 2.3: Overall performance of moral value classifiers on the MFT benchmark datasets. For each dataset, the table reports the F1 scores for the ten MFVs, the non-moral category, and the overall score. The training setting of each model is also indicated: ID, OOD, or in-context. The latter refers to applying the model without fine-tuning on the MFT dataset, typically using zero-shot prompting techniques in LLMs

Congress [168]

Models	**Care**	**Harm**	**Fairness**	**Cheating**	**Authority**	**Subversion**	**Loyalty**	**Betrayal**	**Sanctity**	**Degradation**	**NM**	**Total**	**Setting**
M12 [168]	**0.68**	**0.74**	**0.75**	0.60	**0.69**	**0.79**	**0.64**	**0.70**	0.79	**0.73**	**0.83**	0.72	ID
M13 [168]	0.67	0.73	**0.75**	**0.60**	**0.69**	**0.79**	**0.64**	**0.70**	**0.80**	**0.73**	**0.83**	**0.73**	ID
DAMF-v1 [131]	-	-	-	-	-	-	-	-	-	-	-	0.40	OOD
DAMF-v2 [131]	-	-	-	-	-	-	-	-	-	-	-	0.38	OOD
MFTC-DAMF [131]	-	-	-	-	-	-	-	-	-	-	-	0.38	OOD

Facebook [36]

Models	**Care**	**Harm**	**Fairness**	**Cheating**	**Authority**	**Subversion**	**Loyalty**	**Betrayal**	**Sanctity**	**Degradation**	**NM**	**Total**	**Setting**
MoralBERT [265]	0.64	**0.57**	0.43	0.52	0.48	0.45	**0.58**	0.42	0.57	0.47	-	0.51	OOD
Adv-MoralBERT[265]	**0.65**	**0.57**	0.55	0.42	0.52	0.52	0.56	**0.53**	0.60	0.55	-	0.55	OOD
GPT4 [265]	0.62	0.47	**0.59**	**0.53**	**0.56**	**0.57**	0.56	0.51	**0.65**	**0.59**	-	**0.57**	In-context

Table 2.4: Overall performance of moral value classifiers on the MFT benchmark datasets. The training setting of each model is also indicated: ID, OOD, or in-context. The latter refers to applying the model without fine-tuning on the MFT dataset, typically using zero-shot prompting techniques in LLMs

EMONA [199]

Models	**Total**	**Setting**
Longformer [199]	**0.39**	ID
GPT-4 [199]	0.31	In-context

Covid [163]

Models	**Total**	**Setting**
Congress&MFTC-DAMF [131]	**0.61**	OOD
Congress&eMFD&MFTC-DAMF [131]	**0.61**	OOD
Congress&MFTC-BERT+AFLite [131]	**0.61**	OOD
MFTC-BERT+AFLite [131]	**0.61**	OOD

value classification, and a Weighting Network that estimates linguistic similarity to the target domain and assigns dynamic weights to OOD training instances. Both networks are jointly optimized, enabling the model to emphasize training examples most relevant to the target distribution. Bulla et al. [55] provide a comparative analysis of zero-shot and supervised strategies for cross-domain moral value detection on the MFRC benchmark dataset. Their supervised approach fine-tunes RoBERTa-large separately on each subcorpus and evaluates it on the remaining ones, highlighting robustness and generalization trade-offs across domains. Within supervised cross-domain research, a notable contribution is the Tomea system [211], which offers an explainable framework for analyzing how text classifiers represent moral rhetoric across social domains. Tomea leverages the SHAP method [216], using Shapley values to measure the contribution of each word

to moral predictions and generating domain-specific lexicons that allow for direct comparison of moral cues.

Other relevant studies further explore cross-domain generalization. Liscio et al. [210] evaluate supervised systems for moral value detection using the MFTC benchmark. Extending this to more challenging distribution shifts, Van Luenen et al. [342] demonstrate that a model trained on non-extremist Twitter data can, to a feasible degree, generalize to extremist forum data, albeit with a significant performance reduction. Their comparison reveals that contextual BERT embeddings generalize slightly more effectively than static Word2Vec representations. Complementing this, Nguyen et al. [242] provide a systematic assessment of Transformer-based models, showing that training on multiple domains enhances out-of-distribution generalization compared to single-domain training.

Moral-targeted Language Models Research into moral-targeted LMs extends beyond in-context learning, primarily focusing on two main areas: moral value elicitation via prompt engineering, and the analysis of moral bias embedded within model representations.

Research on moral value elicitation through prompt engineering investigates how strategic prompt design influences model outputs to uncover latent moral tendencies. Abdulhai et al. [1] adapt the MFQ to prompt LLMs like GPT-3 and PaLM, measuring implicit foundation weightings across neutral and adversarial contexts. Their findings indicate that larger models (e.g., Davinci) exhibit moral distributions more closely aligned with human responses, and that intentional prompting can steer perceived political leanings—ultimately influencing downstream behaviors such as charitable donation preferences. In a complementary study, Simmons et al.[325] ex-

amine moral mimicry: the propensity of LMs to replicate partisan moral biases when primed with political identities. Through controlled prompting across multiple datasets (Moral Stories, ETHICS, Social Chemistry), they demonstrate that liberal cues increase reliance on individualizing foundations (Care, Fairness), while conservative cues amplify binding foundation use (Loyalty, Authority, Sanctity). Notably, the magnitude of this bias scales with model size.

Further research probes the critical distinction between abstract moral reasoning and applied judgment in LLMs, often revealing significant inconsistencies. Nunes et al. [245] quantify moral hypocrisy by prompting models like GPT-4 and Claude 2.1 with both the MFQ and MFV. They find that while models exhibit human-comparable consistency within each dataset, they display systematic contradictions across them, highlighting a disconnect between stated values and contextualized decisions. He et al. [139] evaluate behavioral alignment. They introduce an affective alignment metric to compare LLM-generated ideological tweets to human-authored content on contentious issues. Their analysis across 36 LLMs reveals a systemic liberal affective bias and demonstrates that the affective misalignment of models extends beyond existing human partisan divides.

Prompt-based moral analysis has also been applied in persuasion and social media studies. Carrasco-Farre [66] adapts prompts from persuasion experiments to elicit arguments on 56 claims, showing that LLMs match human persuasiveness, use more complex structures, and employ moral framing more frequently. Jiang et al. [162] introduce Social-LLM, which implicitly incorporates prompt-driven content encoding with social network features, achieving competitive morality-related detection across large-scale Twitter datasets.

The second research direction focuses on the analysis of moral bias in LM embeddings. Kennedy et al.[174] demonstrate that BERT embeddings capture linguistic differences in moral concerns more effectively than lexicon-based measures. Xie et al.[357] develop text-based methods to predict human moral judgments, showing that S-BERT embeddings achieve the highest accuracy. Hämmerl et al. [135] examine cross-lingual transfer of moral standards, highlighting how norms learned from dominant languages (e.g. English) tend to be imposed on others, with potentially negative consequences. Their results indicate that models struggle with complex sentences and negation, and that performance varies depending on language and training data size.

2.1.2 Basic Human Value

The theory of Basic Human Values (BHV) [311] conceptualizes values as universal motivational goals that shape human decisions and behaviours across cultures. These values are theorized to arise from the interaction of three fundamental human requirements: the satisfaction of individual biological needs, the demands of social coordination, and the imperatives of group survival and well-being. Although values serve as relatively stable guiding principles, their prominence and influence can vary across cultural settings and individual circumstances.

In its original formulation, the BHV identifies ten basic value types: self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence, and universalism. These values are organised in a circumplex structure that reflects both motivational compatibility and potential conflict: openness to change contrasts with conservation, while self-enhancement is opposed to self-transcendence (Figure 2.2). A later extension

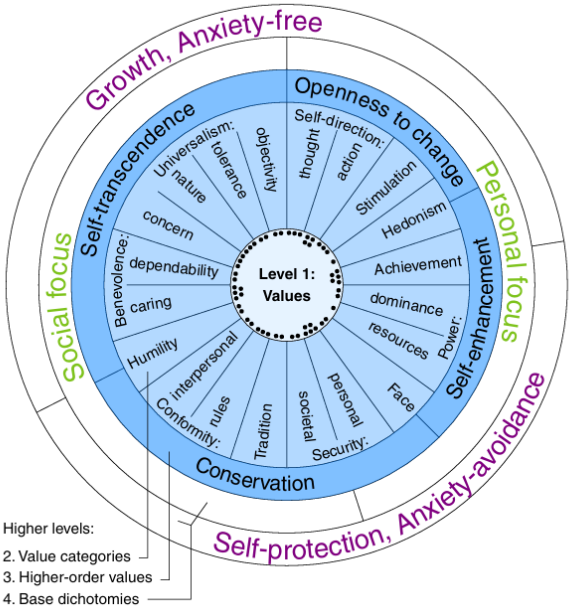


Figure 2.2: Basic Human Values (BHV) theory framework. At the first level, the taxonomy comprises 54 values (depicted as black dots), which are organized across the more abstract Levels 2-4. In the BHV value wheel, categories that tend to conflict are positioned on opposite sides

of the model subdivides these categories into nineteen more fine-grained values, thereby preserving cross-cultural validity while enabling more nuanced analyses of individual and collective value orientations.

Section 2.1.2 introduces the benchmark datasets developed within this theoretical framework, which provide the foundation for evaluating AI-based moral value classifiers. Section 2.1.2 presents state-of-the-art models across two methodological paradigms: ID and OOD evaluation. We further describe recent approaches that leverage LLMs as zero-shot, unsupervised classifiers of moral values, employing different prompting strategies.

Dataset

Among the benchmark datasets grounded in BHV theory, Kiesel et al.[178, 179], Qiu et al.[272], Yao et al.[363], and Borenstein et al.[49] present main contributions (cf. Table 2.5).

Kiesel et al. [178] introduce the Webis-ArgValues-22 dataset, which comprises approximately 5,000 arguments collected from social media and annotated by human annotators according to Schwartz's value theory. Building on this foundation, the SemEval-2023 ValueEval shares task [179] extended the dataset and framed the problem as a multi-label classification challenge aimed at detecting implicit values in arguments. The resulting corpus, referred to as Touché23-ValueEval, comprises more than 9,000 annotated arguments, each associated with one or more categories derived from the BHV taxonomy. Whereas the Webis-ArgValues annotation scheme distinguishes four hierarchical levels of the BHV taxonomy, corresponding to the structure of the BHV value wheel (cf. Figure 2.2), the Touché23-ValueEval dataset restricts its scope to the two inner levels of this taxonomy. These levels comprise 20 and 54 labels, respectively. To ensure reliability, multiple annotators label each argument independently, with the majority (89%) of instances assigned to more than two value categories.

Qiu et al. [272] introduce the ValueNet dataset, which contains more than 21,000 textual scenarios annotated with respect to ten basic human values (i.e. Self-Direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence, and Universalism) and further grouped into four higher-order categories: Openness to Change, Self-Enhancement, Conservation, and Self-Transcendence. Each scenario is represented as a ten-dimensional utility vector, encoding positive, negative, or neutral orientations toward the corresponding values, with annotations provided by

multiple crowdworkers.

FULCRA [363] corpus extends this line of research to the evaluation and alignment of LLMs. The dataset consists of 5,000 prompt–response pairs annotated with 58 fine-grained value elements, which are subsequently mapped onto the ten basic and four higher-order values defined in Schwartz's theory. Each instance is encoded as a discrete stance vector, indicating whether a response aligns with, opposes, or is irrelevant to a given value. Annotation is carried out through a hybrid pipeline: GPT-4 generated initial predictions, which are then refined by expert human annotators, particularly in cases of low agreement between the model and humans. This methodology highlights the potential of combining automated and expert annotation for advancing value alignment research.

Borenstein et al. [49] conduct a large-scale computational analysis of Reddit to examine the distribution and expression of values in user-generated content. The dataset encompasses more than nine million posts from over eleven thousand subreddits, which are processed using automated classifiers to estimate value-relevance probabilities for each of the ten basic values and, when relevance is high, to further predict stance as positive, negative, or neutral. Human evaluations on sampled posts assess both relevance and stance, achieving moderate-to-high inter-annotator agreement and confirming that classifier predictions with high confidence were generally reliable.

Methods and Results

Current research on identifying human values in text based on Schwartz's BHV theory predominantly relies on Transformer-based architectures, encompassing both encoder-decoder and decoder models [178, 179, 272, 307, 363, 377] (cf. Table 2.6).

Table 2.5: Overview of benchmark datasets for moral value detection within the BHV framework. For each corpus, we report its size, the level of the BHV taxonomy considered (i.e. taxonomy level), and the number of annotators

Dataset	**Size**	**Taxonomy Level**	**Annotator**
Touche23-ValueEval [179]	9,324	2	3
Webis-ArgValues-22 [178]	5,270	4	3
ValueNET [272]	21,374	1	4
FULCRA [363]	5,000	1	3
Reddit [49]	12,000	1	3

Qiu et al. [272] develop a regression framework in which BERT- and BART-based Transformer models are fine-tuned to predict continuous utility scores for social scenarios along the second level of the BHV taxonomy, encompassing ten moral dimensions. Specifically, the input text is concatenated with a special token indicating the target value (e.g., [CLS][$VALUE]), and a regression head outputs a continuous score between –1 and +1, where –1 denotes opposition, +1 indicates alignment, and 0 represents neutrality. They evaluate their model on the ValueNet dataset, achieving an accuracy of 0.64 and an F1 score of 0.57 in a ID scenario.

Similarly, in the ValueEval 2023 shared task [179], most participants employ fine-tuned Transformer-based models, often in ensemble configurations. The winning system, Adam Smith [307], combines 12 models (primarily DeBERTa- and RoBERTa-based models) by averaging their predictions and optimizing a global threshold on a held-out validation set.

Yao et al. [363] fine-tune language models as binary classifiers, experimenting with BERT-, DeBERTa-, and LLaMA-2-7b-based architectures. Their method incorporates BHV definitions directly into the input as prompt response pairs and performs classification iteratively across all ten value dimensions to construct complete value vectors. Yao et al. evaluate this

Table 2.6: Overall performance of moral value classifiers on the BHV benchmark datasets. For each dataset, the table reports the overall accuracy and F1 scores. The training setting of each model is also indicated: ID, OOD, or in-context. The latter refers to applying the model without fine-tuning on the BHV dataset, typically using zero-shot prompting techniques in LLMs. Bold values denote the best results among models trained and evaluated on the same distribution (ID and ID + in-context), as these settings are directly comparable due to their alignment between training and test data distributions.

Touche23-ValueEval [179]

Model	**Accuracy**	**F1**	**Setting**
EAVIT [377]	0.94	**0.69**	ID + In-context
Adam Smith [307]	-	0.57	ID
CoT-GPT4 [377]	0.89	0.58	In-context

Webis-ArgValues-22 [178]

Model	**Accuracy**	**F1**	**Setting**
EAVIT [377]	0.92	0.66	ID + In-context
CoT-GPT4 [377]	0.86	0.56	In-context

FULCRA [363]

Model	**Accuracy**	**F1**	**Setting**
DeBERTa-large [363]	0.82	-	ID
LlaMa-2 7B [363]	**0.83**	-	ID

ValueNET [272]

Model	**Accuracy**	**F1**	**Setting**
BART [272]	0.64	0.57	ID
EAVIT [377]	**0.78**	-	ID + In-context
GPT-4o-mini [377]	0.78	-	In-context

approach on the FULCRA dataset, achieving an accuracy score of 0.82 and 0.83 for the DeBERTa-large and LLaMA-2 models, respectively.

More recently, Zhu et al. [377] introduced EAVIT, a hybrid framework that combines fine-tuning with in-context learning techniques. The system first applies a value detector, based on models such as LLaMA-2-13b-chat fine-tuned with QLoRA and explanation-based training, to filter out clearly relevant or irrelevant values and generate candidate sets for ambiguous cases. In the final stage, online LLMs such as GPT-4 and GPT-4o are prompted selectively using only the candidate set definitions. This selective prompting strategy reduces context length and computational cost while preserving high accuracy in value identification. For evaluation, EAVIT was benchmarked against both traditional NLP models (e.g., BERT, RoBERTa, and the SemEval-2023 Task 4 winner Adam Smith [307]) and a broad set of LLM baselines, including GPT-2, LLaMA-2, GPT-3.5, GPT-4o-mini, and GPT-4, tested under diverse prompting strategies (single-step, multi-step, sequential, and Chain-of-Thought). While unsupervised zero-shot and few-shot approaches provided meaningful reference points, EAVIT consistently outperformed all baselines across three benchmark datasets, i.e. Webis-ArgValues-22 [178], Touché23-ValueEval [179], and ValueNet [272]. Notably, it surpasses GPT-4 in the sequential Chain-of-Thought setting, achieving state-of-the-art results with accuracies and F1 score of 0.94 and 0.69 on Touché23-ValueEval and 0.92 and 0.66 on Webis-ArgValues-22, respectively. On ValueNet dataset, EAVIT presents comparable performance w.r.t. GPT-4o-mini baseline with an accuracy of 0.78.

Beyond datasets and systems explicitly targeting BHV-based value identification, Schwartz's theory has been employed as a conceptual anchor in broader applications and evaluations. Obie et al.[247] analyzed mo-

bile app reviews to detect value violations, while Fischer et al.[108] show that ChatGPT reproduces many distinctions from the Portrait Value Questionnaire [310, 312], though with partial merging of socially oriented values. Complementary resources such as CLAVE [364] and ValueBench [280] integrate BHV into value-sensitive evaluations of generative systems, and methodological contributions include adaptive benchmarks like AdAEM [95] and alignment techniques such as ConVA [166]. Interdisciplinary efforts, e.g., ValueLex [41], further interrogate whether human-centered theories like BHV suffice to capture value orientations in LLMs. Collectively, these works demonstrate the growing role of BHV not only as an annotation scheme but also as a reference framework for evaluating and aligning AI systems with human moral expectations.

2.1.3 Morality-as-Cooperation

The Morality-as-Cooperation theory [78, 79] argues that morality is not a single, unified construct but rather a collection of biological and cultural adaptations evolved to address the recurring challenges of cooperation in human social life. Grounded in the mathematics of cooperation and evolutionary game theory, this framework identifies seven distinct classes of cooperative dilemmas and their corresponding adaptive solutions, which together constitute the fundamental "elements" of human morality. These domains include: family values (arising from kin selection), group loyalty (from mutualism and coordination), reciprocity (from social exchange), heroism (hawk-like strategies in conflict resolution), deference (dove-like strategies), fairness (principles of resource allocation), and property rights (the recognition of possession). Importantly, the MAC theory conceptualizes morality as a combinatorial system, in which seven basic elements can be combined

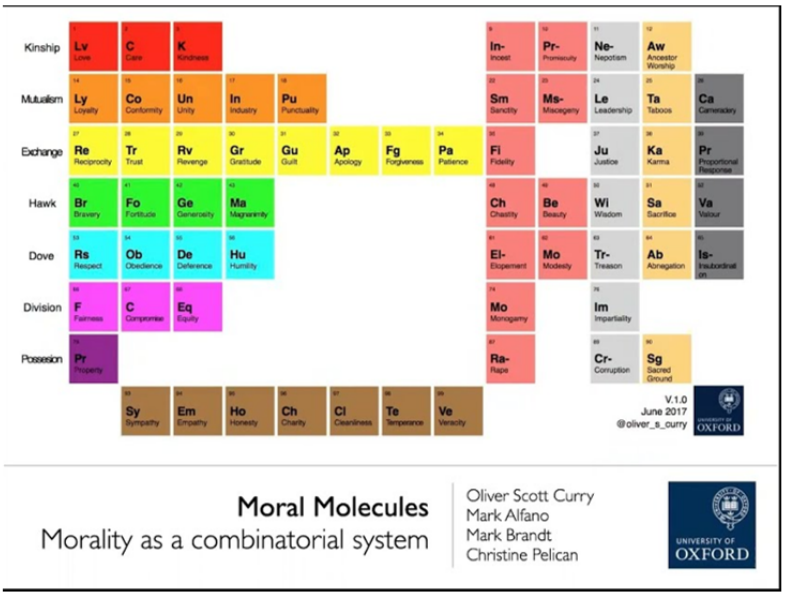


Figure 2.3: Morality-as-cooperation framework

to produce a wide variety of more complex moral "molecules". Morality-as-Cooperation theory provides a principled and predictive taxonomy, described as a "Periodic Table of Ethics" (Figures 2.3), intended to capture both the structure and content of actual and possible moral systems.

The application of MAC theory within NLP remains relatively underexplored. Existing contributions include the works of Malik et al.[220], Alfano et al.[10], and Söderholm et al.[327], which are summarized in Table2.7.

Malik et al. [220] introduce the Extended Morality-as-Cooperation Dictionary (eMACD), derived from crowd-sourced annotations of over 56,000 text spans from news articles. The dataset comprises 56,125 annotated highlights extracted from 3,342 unique articles, covering 18,619 unique words. Annotations were provided by 1,070 U.S.-based contributors recruited via Prolific Academic, with each span linked to one of the seven MAC domains (i.e. family values, group loyalty, reciprocity, heroism, deference, fairness,

and property rights), further categorized into virtue and vice. Each lexical item is represented as a probability vector across the seven domains. The authors assess eMACD's predictive validity with machine learning models, including random forest classifiers and linear regression, across ten validation studies spanning news, social media, and film dialogue corpora. Results indicate that eMACD consistently outperforms existing moral dictionaries.

Alfano et al. [10] construct the Human Relations Area Files (HRAF) corpus by applying the MAC-D lexicon to 9,653 paragraphs from 1,389 documents representing 256 societies. Söderholm et al. [327] examine the applicability of MAC to ancient texts, annotating 102 units from the Sermon on the Mount in Koine Greek. While their analysis confirms broad coverage of MAC domains, it also reveals substantial annotator divergence, underscoring the challenges of preparing reliable training material for machine learning applications.

Table 2.7: Overview of benchmark datasets for moral value detection within the MAC framework. We report its size, the taxonomy size, and the number of annotators for each dataset.

Dataset	**Size**	**Taxonomy**	**Annotator**
eMACD [220]	3,342	7	1,070
HRAF [10]	9,653	7	-
Sermon [327]	102	7	5

2.2 Towards Assessable and Controllable AI in Human-centered domain

The development of AI systems that can operate reliably in morally sensitive contexts demands more than strong predictive accuracy; it also requires robustness, interpretability, and controllability. This section addresses two

complementary dimensions that are central to advancing Human-Centered AI in such domains. Section 2.2.1 considers the challenges posed by subjective annotation, introducing metrics and evaluation strategies that more faithfully capture human diversity in relation to model performance. Section 2.2.2 explores controllability, with a focus on grammar-constrained decoding (GCD) methods designed to enforce structural compliance in generative models. Collectively, these perspectives provide a methodological foundation for developing AI systems that are both resilient to real-world variability and aligned with human values.

2.2.1 Evaluating Subjective Labels: Annotator Diversity, Agreement, and Model Alignment

Recent research has increasingly focused on addressing the limitations of supervised methods, particularly in scenarios where training data are scarce or annotations are highly subjective (cf. Sect. 1.3.2). Moral value detection exemplifies these challenges, as it often relies on judgments that are inherently subjective, leading to limited datasets and potential annotation biases (cf. Sect. 1.2.2). To mitigate these issues, Golazizian et al. [123] introduced a two-stage framework designed to improve the efficiency of annotation collection and modeling. In the first stage, a small group of annotators is employed to construct a multitask model; in the second stage, a limited number of samples are selected and annotated per annotator, thereby integrating their unique perspectives into the model. To support evaluation at scale, the authors released the Moral Foundations Subjective Corpus, a dataset comprising 2,000 Reddit posts annotated by 24 individuals for moral sentiment. Their findings show that this framework outperforms previous state-of-the-art methods in capturing individual annotator perspectives while requiring

only 25% of the typical annotation budget. Moreover, the approach fosters fairness by reducing disparities in performance across annotators. In a complementary line of work, van der Meer et al. [341] proposed Annotator-Centric Active Learning (ACAL), a framework that combines data sampling with annotator selection strategies. ACAL pursues two main objectives: first, to efficiently capture the diversity of human judgments in subjective tasks, and second, to evaluate models using annotator-centric metrics that prioritize minority perspectives alongside majority ones. Through experiments spanning seven subjective NLP tasks, the authors tested a range of annotator selection strategies, adopting both established and novel human-centered evaluation measures. Their results demonstrate that ACAL improves data efficiency and excels in capturing diverse perspectives, though its effectiveness depends on access to a sufficiently large and heterogeneous pool of annotators. Our research adopts a distinct direction by shifting from supervised approaches to unsupervised learning with LLMs. This eliminates reliance on pre-labeled data, enhancing scalability and adaptability in real-world contexts where annotations are limited or costly. In addition, we introduce a validation framework that moves beyond traditional evaluation strategies by assessing LLM performance in the moral domain without supervised labels. This approach offers a new perspective on the evaluation of models in highly subjective tasks. The evaluation of machine learning models has traditionally relied on well-established performance metrics such as precision, recall, and F1-score, which balance false positives and false negatives [87]. These metrics, however, are not directly applicable to human annotations, which are typically evaluated using inter-annotator agreement measures such as Cohen's Kappa [344], Fleiss's Kappa [104, 110], and Krippendorff's alpha [193]. Cohen's Kappa accounts for chance agreement but

is restricted to two annotators and is thus unsuitable for more complex settings. Fleiss's Kappa extends this to multiple annotators by estimating an overall agreement score, though it fails to capture fine-grained distributional differences. Krippendorff's alpha offers greater flexibility, handling multiple data types including ordinal and interval scales, but like the other measures, it assumes a fixed ground truth and primarily assesses consistency rather than performance relative to a reference standard. These assumptions limit their applicability in tasks where subjectivity and ambiguity are inherent, resulting in an incomplete representation of annotation reliability [275]. Alternative measures have been proposed to address some of these shortcomings. The pairwise F1-score [154], for instance, evaluates agreement between annotators at the instance level, capturing finer-grained annotation differences by focusing on label overlap rather than binary matches or mismatches. While this method better reflects subtle divergences, it remains limited to pairwise comparisons, does not generalize to variable numbers of annotators, and fails to account for chance agreement. Consequently, distinguishing genuine agreement from random overlap remains problematic. Our proposed metric builds on these principles, aiming to provide a more holistic framework that integrates inter-annotator variability while maintaining closer alignment with established model evaluation methodologies. The complexity of subjective annotation becomes even more evident in multi-class and multi-label tasks, where inconsistencies often arise from differing interpretations of class boundaries or varying degrees of granularity [224]. For instance, Yang et al. [362] highlighted the limitations of traditional agreement measures in contexts characterized by high inter-observer variability, such as segmentation tasks. To address these challenges, researchers have explored consensus-based approaches, such as crowdsourcing frameworks [97, 98, 122, 270] and

probabilistic annotation techniques [263, 376]. Probabilistic models, for example, attach confidence scores to labels rather than treating them as discrete categories, while consensus-based methods aggregate multiple annotator judgments to construct more reliable gold standards. Although these approaches improve annotation quality, they remain focused on optimizing label reliability rather than aligning human annotations directly with model evaluation. Despite such advancements, existing methodologies continue to struggle with integrating human annotation and model evaluation within a unified framework. Consensus-driven strategies privilege agreement, often obscuring meaningful individual variation, while active learning methods improve data efficiency but fall short of addressing the core challenge of interpretability in the comparison between models and annotators. In Chapter 4, we propose a novel methodology that seeks to bridge this gap by introducing an evaluation metric that explicitly links annotation quality with model performance, thereby providing a more interpretable and robust framework for the assessment of complex annotation tasks.

2.2.2 Grammar Constraints and Decoding Strategies

In the broader effort to develop AI systems that are more controllable and therefore safer, a central goal of HCAI (cf. Sect. 1.2.3), Grammar-Constrained Decoding (GCD) has emerged as one of the most actively explored research directions. This line of work has gained traction in the current LLM-dominated landscape, where controllability has become a critical challenge. As discussed in Sect.2.1, unsupervised and semi-supervised approaches are often preferred over supervised ones, as they support stronger generalization and better leverage modern generative models. However, this shift limits the role of explicit supervision and fine-tuning, increasing the risk

of overspecialization in highly subjective domains. Instead, such methods depend on implicit knowledge, restricting the enforcement of strict output conformity. As a result, generative models frequently exhibit issues such as hallucinations [157], undermining their reliability in sensitive contexts.

To mitigate these risks and enforce predefined constraints, several strategies have been proposed. Prompt-based few-shot techniques attempt to guide models toward producing rule-abiding outputs while preserving generalization, but they remain inherently unreliable, as no guarantee exists that outputs will adhere to the intended rules. In response, Grammar-Constrained Decoding has recently emerged as a promising and more principled alternative.

A number of GCD frameworks [88, 120, 191, 204, 260, 261, 306, 318, 353] have been proposed to enforce structured output constraints by integrating formal grammatical mechanisms into the decoding process. These approaches rely primarily on finite-state automata (FSA), context-free grammars (CFGs), and deterministic parsing strategies. Koo et al. [191] and Park et al. [261] exploit automata-based techniques to enforce structural compliance while improving computational efficiency. Koo et al. employ finite-state transducers (FSTs) and finite-state machines (FSMs), offering closed-form solutions for regular languages and extending the approach to deterministic context-free languages through pushdown automata (PDAs). Their method provides substantial speedups in constraint enforcement and allows modular adaptation across structured generation tasks. Park et al. refine this strategy by introducing a token spanner table that efficiently maps LLM tokens to sequences of grammar terminals, thus optimizing both offline preprocessing and online constrained decoding. Li et al. [204] present Formal-LLM, which constrains plan generation using

non-deterministic PDAs. Their method allows PDAs to be constructed directly from natural language, with transitions integrated into LLM prompts to guide text generation, though it does not intervene during the decoding process itself. In contrast, Willard and Louf [353] introduce Efficient Guided Generation, modeling text generation as a sequence of FSM state transitions. This enables constraint enforcement with $O(1)$ complexity per token, but restricts applicability to regular languages, limiting its scope for more complex structures. Other methods, including Geng et al. [120], PICARD [306], and Shin et al. [318], adopt grammar-based decoding strategies built upon input-dependent grammars (IDGs) or external parsing mechanisms, which are especially suited to structured tasks such as semantic parsing. Park et al. [260] propose Grammar-Aligned Decoding (GAD), designed to preserve the LLM distribution over valid outputs while building on the formal framework of Geng et al. [120], making it complementary to existing grammar-based methods. Along similar lines, Ahmed et al. [5] employ logic circuits and locally constrained resampling to enforce structural compliance while maintaining expressiveness, correcting biased samples through importance weighting and resampling. Zhang et al. [370, 371] integrate Hidden Markov Models (HMMs) to impose logical constraints on LLM outputs, using probabilistic reasoning to evaluate token validity and compute conditional probabilities during generation. Unlike FSM-based methods, their approach supports a broader class of context-free grammars. Unlike probabilistic or external parsing methods, it guarantees strict adherence to grammar rules without resorting to superlinear processing or post-filtering, thereby offering greater scalability. Deutsch et al. [88] propose a general-purpose algorithm for constrained sequential inference, integrating constraints into the decoding process through deterministic pushdown automata. Their ap-

proach treats the automaton as an input specification but does not define which classes of grammars can be handled efficiently. Dong et al. [93] extend this line with XGrammar, a structured generation engine for LLMs that enables flexible constrained decoding based on context-free grammars. By partitioning the vocabulary into context-independent tokens, XGrammar reduces the computational overhead usually associated with grammar-based decoding. However, it does not guarantee linear-time efficiency, as it relies on non-deterministic PDAs to resolve ambiguities, requiring multiple concurrent stacks, which may become computationally intractable.

In Chapter 5, we introduce GRAMMAR-LLM, a system that overcomes these limitations by enforcing constraints directly during decoding via deterministic pushdown automata derived from LL(prefix) grammars. This guarantees that generated outputs strictly adhere to predefined syntactic structures while ensuring that computational overhead grows linearly with the length of the generated text. Unlike more general approaches, GRAMMAR-LLM thus combines scalability with strict structural compliance, making it a robust and adaptable solution for constrained text generation.

2.3 AI for the Deaf Community

One of the central goals of HCAI is to design AI systems that promote inclusivity (cf. Sect. 1.3.1). Within this scope, a particularly relevant application in NLP is the development of translation systems between spoken or written language and sign language. Sign Language Translation (SLT) has recently emerged as a rapidly expanding frontier of AI research. However, current generative models such as GPT are ill-suited for this task, as sign language data are largely absent from pre-training corpora. Consequently, conventional prompting-based translation techniques remain inadequate for

SLT.

Chapter 8 presents a novel contribution to the field by introducing a new sign language (SL) translation model for NLP that leverages the capabilities of current LLMs to perform the task. To provide background and a literature overview, Section 2.3.1 reviews the foundations of SL translation, while Section 2.3.2 details the current state of the art for this task.

2.3.1 Theoretical foundations of Sign Language translation

Sign languages are fully developed natural languages that employ a complex system of manual gestures, facial expressions, and body movements to convey meaning. Unlike spoken languages, they are inherently visual-spatial, relying on three-dimensional space to express grammatical and semantic information. This unique modality poses significant challenges for linguistic and computational analysis. Intermediate representations, such as gloss notation, HamNoSys (Hamburg Notation System) [266] and SignWriting [333], have been developed to bridge these gaps. Glosses provide written approximations of sign language using natural language words or symbols to label individual signs, simplifying transcription and analysis. HamNoSys, on the other hand, offers a phonetic transcription system that encodes the physical parameters of signs, including handshape, orientation, movement, and location, in a highly abstract and language-agnostic manner. While these systems are effective for specific applications, they often fail to capture the holistic visual-spatial structure of sign languages. In contrast, SignWriting provides a visually expressive and iconic featural system explicitly designed to capture the intricacies of signed communication. It represents signs graphically, encoding key features like handshapes, orientations, movements, body locations, and facial expressions within a two-dimensional "sign box" that

mirrors the structure of signed expressions. Additionally, SignWriting can be linearized outside the sign box, enabling its integration into practical applications similar to written scripts for spoken languages. This dual representation – spatial and linear – positions SignWriting as a powerful tool for linguistic analysis and computational applications, such as sign language recognition [4], and translation. SignWriting has two primary computerized specifications: Formal SignWriting in ASCII (FSW) and SignWriting in Unicode (SWU). Both specifications are written linearly, aligning with the conventions of written languages. Each FSW sequence of symbols begins with a bounding box marker (M), followed by positional factors (x, y) and symbol identifiers. These identifiers include a base character (e.g., S1ce) along with modifiers for orientation, rotation, and spatial metadata. Figure 2.4 provides an example of FSW encoding, demonstrating the systematic decomposition and representation of signs. FSW adheres to a rigorous syntax defined by regular expressions, guaranteeing the grammatical and syntactical correctness of encoded signs. Moreover, FSW is isomorphic with SWU enabling seamless interchange between formats. By encoding both the symbolic and spatial features of signs, FSW notation enables the decomposition of signs into constituent components (e.g., handshapes, movements) and leverages positional data for semantic modeling. This factorization facilitates the separation of visual features from semantic features and therefore provides a transparent and interpretable representation, which in turn aids in developing robust and reliable translation systems. In addition to the FSW representation, a sign can be paired with a glossing system specifically tailored to SignWriting. These notations, referred to as canonical descriptions, offer a standardized and abstract representation of sign language, effectively bridging the gap between its visual-spatial nature and linguistic

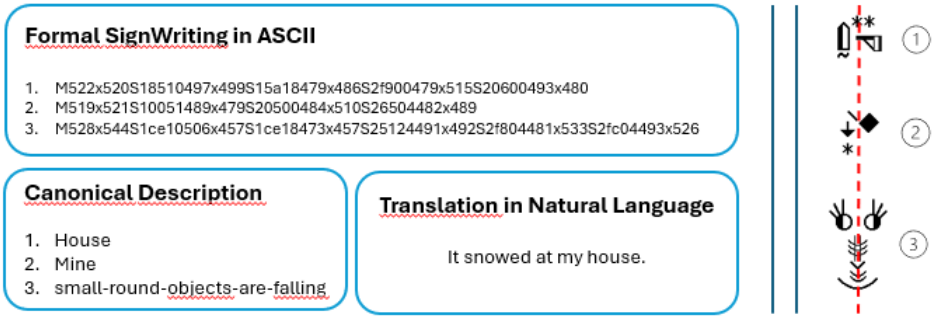


Figure 2.4: Example of FSW encoding and canonical descriptions of a sign language sequence. FSW provides a detailed, structured representation of signs, while canonical descriptions offer a more abstract and language-independent representation.

modalities. By converting complex sign language gestures into structured natural language descriptions, canonical descriptions facilitate efficient machine processing and analysis.

In Section 8, we present a novel contribution that uses FSW and canonical descriptions to investigate translation tasks, demonstrating their efficacy in both data-rich and data-scarce settings, and their compatibility with contemporary computational frameworks.

2.3.2 Sign Language Translation: an Overview

Recent advancements in SLT research primarily focus on two categories of models: end-to-end models, which directly map video inputs to text outputs, and representation-based models, which employ intermediate symbolic representations such as glosses or other meta-languages to decompose the task into smaller and interpretable steps [85, 105]. While end-to-end models offer efficiency [375], gloss-based models remain widely adopted for their interpretability and modularity. These multimodal models are partic-

ularly well-suited for video-to-text SL translation, whereas the inverse task of text-to-sign translation remains substantially more challenging. Generating temporally coherent and linguistically accurate sign language videos is still a complex endeavor, and current solutions largely rely on animation or avatar-based methods [367]. In this case, intermediate symbolic representations play a key role, as they reduce the complexity of video generation and can be integrated with downstream avatar animation tools.

State-of-the-art approaches to text-based SLT, which aim to address both SL-to-text and text-to-SL tasks without relying on video data, typically adopt structured intermediate representations such as glosses [20, 366], HamNoSys [173], SignWriting, or SMPL-X. Glosses provide simple written approximations of signs using natural language labels, but they lack the formalism required to encode phonological, spatial, and non-manual features, which limits their usefulness for synthesis or generation. Similarly, HamNoSys provides a phonetic transcription system for sign languages, but it lacks a standardized computational form suitable for direct integration into neural models. By contrast, representations such as SignWriting and SMPL-X have been explicitly designed to encode visual-linguistic properties, making them more suitable for generation pipelines. Several studies have explored the use of SignWriting for recognition and segmentation [238, 315], with advances in languages such as Arabic and Japanese [11, 227], as well as ontology-driven approaches [12].

Recent translation approaches increasingly rely on transformer architectures [84, 164]. The work of Jiang et al. [164] is particularly relevant, as it establishes the state of the art in SignWriting-based SLT. They propose a multilingual pipeline that translates natural language into SignWriting by decomposing the process into multiple stages, including analysis, factoriza-

tion, decoding, and evaluation of SignWriting sequences. Their system represents SignWriting sequences as sequences of Formal SignWriting (FSW), decomposed into base units (symbols) and supplementary information (e.g., positional numbers). These elements are encoded as variables processed by a factorized model [116, 186]. For spoken-to-sign tasks, symbol tokens are decoded using Beam Search, while positional factors are predicted through regression. To further improve performance, external dictionary resources such as Dictionary Puddles[1] are integrated. Despite these advances, the model exhibits limitations: finger-spelling requires specialized handling, and conventional text classification metrics may be insufficient for evaluating SLT outputs, given their inherently visual and spatial characteristics.

More recently, LLMs have been applied to SLT, particularly in video-to-sign translation [124, 159, 165, 208, 356]. These methods have been explored both with and without intermediate representations such as gloss notation [170, 213]. In addition, LLMs have shown potential in translating between natural language and glosses [106, 196]. However, to the best of our knowledge, no study has directly addressed the task of sign language translation from natural language to encoded symbolic representations such as SignWriting using generative pre-trained LLMs. This limitation remains a key challenge, since LLMs have not been exposed to sign language representations during pre-training, raising fundamental questions about their capacity to handle such tasks.

[1] https://www.signbank.org/signpuddle/

3 Moral Value Detection

In Chapters 1 and 2, I detailed how advancing HCAI requires systems capable of recognizing and operationalizing moral values in natural language processing. However, identifying moral values in textual documents is a persistent challenge, mainly due to the limitations of current supervised models. These models often struggle with overfitting problems, as they over-conform the specific training data distribution [56].

Although these classifiers show good results for datasets with similar distributions, their performance suffers statistically significant degradation when dealing with data from a real scenario with an unknown data distribution (cf. Chapter 2.1). This contrast underscores the inherent trade-off between supervised approaches, which can be highly effective but remain tied to domain-specific training data, and zero-shot methods, which leverage large pre-trained models to perform classification without task-specific fine-tuning. The latter offer the advantage of greater adaptability, as they can generalize more effectively across unseen or heterogeneous domains.

This highlights the need for more versatile approaches to moral classification. Researchers can achieve this by addressing the limitations of domain-specific training data and implementing methods that adapt to different data distributions. Such advances promise to impact different areas, including content moderation, ethical decision-making frameworks, and the

broader landscape of human-computer interaction.

A further key aspect of moral detection concerns the evaluation methodology (cf. Sect. 1.3.2, Sect. 2.2.2). Current gold standards rely on annotations from multiple human coders, but this approach faces limitations due to low inter-annotator agreement for inherently subjective tasks. This inconsistency weakens the gold standard's reliability, hindering accurate assessment of model performance.

In addressing these challenges, this Chapter presents the contribution [59], which leverage the ability of state-of-the-art models to grasp abstract social concepts through training on extensive pools of common-sense data [24]. This strategy aims to mitigate the risk of overfitting specific moral datasets and develop models that generalize across real-world scenarios. To validate our approach, we present an in-depth evaluation comparing the performance of state-of-the-art unsupervised models with human judgments. Our novel framework directly juxtaposes model outputs with human annotations. This head-to-head approach addresses subjectivity concerns by placing human and model ratings on equal footing, offering valuable insights into both.

Our methodology explores the use of state-of-the-art LLMs of different sizes (i.e. Mistral-7B, Llama-7B, Mixtral-8x7B, Llama-70B, GPT-3.5, and GPT-4) as zero-shot, pre-trained unsupervised multi-label classifiers for moral value detection. This approach allows us to identify moral values without the need for explicit training on the annotated data. To evaluate the technique's adaptability and effectiveness, we incorporate both zero-shot and few-shot settings. The latter uses a limited number of examples to guide predictions, enhancing the models' contextual understanding and performance. To evaluate the effectiveness and adaptability of our technique, we compare it with a smaller zero-shot method based on Natural Language

Inference (NLI). Our assessment reveals GPT-4 as the top performer, followed by GPT-3.5, Llama-70B, Mixtral-8x7, Mistral-7B and Llama7B. Notably, all systems outperform the NLI-based model. Furthermore, we delve into the strengths and limitations of the models in different value domains and prompting configurations. Our exploration includes everyday morality and morality applied to political contexts, employing two distinct prompting approaches that leverage the models' contextual understanding capabilities.

For a comprehensive and unbiased evaluation, we design a strategy that divides moral detection tasks into two sub-tasks: (i) multi-label moral value detection and (ii) binary moral classification. The multi-label moral value detection subtask focuses on identifying moral values in a text, assuming the model recognizes the text as inherently moral. The binary classification sub-task determines whether a text contains moral content or not. This strategy allows for a nuanced and in-depth understanding of findings in the field of moral classification. Through separate analyses of these two aspects, we aim to provide a clearer perspective on model capabilities. This ensures that the evaluation of the simpler binary task does not overshadow the more complex multi-label task. Furthermore, we introduce a novel experimental framework that allows a direct comparison between human and model performance under identical conditions. This comparative analysis juxtaposes results from automated value detection with those derived from human evaluation, clarifying the moral dimensions in which models achieve superior or comparable performance to humans.

In our research, we employ the MFRC [337], comprising 16,123 Reddit comments. This corpus is divided into three separate subcorpora, each delineating distinct domains (cf. Sect. 2.1.1). Furthermore, annotations in the corpus adhere to Graham and Haidt's MFT [129].

The Chapter is structured as follows. Section 3.1 delves into the unsupervised methodologies we use, distinguishing between the techniques involving LLM (Sect. 3.1.1 and the NLI-based method (Sect. 3.1.2. In Section 3.2, we present a comprehensive overview of our experimental setting, exploring an assessment of different linguistic models in the task of moral value detection (Sect. 3.2.2. We specifically focus on two key sub-tasks: multi-label identification of moral value dimensions (Sect. 3.2.2 and binary detection of text containing moral content (Sect. 3.2.2. Additionally, in Section 3.2.3, we introduce a direct human model comparison framework for evaluating both multi-label and binary moral detection that addresses the issue of subjectivity limitations in gold standards. Furthermore, Section 3.3 provides a detailed discussion of the results obtained across all task configurations and evaluation settings [1].

For a comprehensive discussion of the theoretical foundations and the current state of the art, see Chapter 2.1.1.

3.1 Methodology

We leverage a set of pre-trained LLMs and an NLI model as zero-shot ready-made unsupervised multi-label moral value classifiers. Our goal is to evaluate their effectiveness in identifying moral content in textual data. Section 3.1.1 provides an overview of the LLMs we employ, including a comprehensive analysis of the prompting techniques we use. Section 3.1.2 delves into the architecture of the NLI-based model detailing the different configurations we implement.

[1] All experiment results and the corresponding code for replication are available at the following link: $https://osf.io/kw6rc/?view_only = 73c3d8e2f7c9483ea012347d70d27b6f$

3.1.1 LLM-based models

We employ GPT-4 [2], GPT-3.5 [52, 252][2], Llama3-70B [231], Mistral-7B [161], and Mixtral-8x7B[3] and Llama2-7B [336] models in unsupervised zero-shot and few-shot settings to assess LLMs' capability in detecting moral values in text. GPT-3.5 is specifically designed to excel at interpreting and executing instructions, aiming to deliver consistent and contextually relevant responses. Its training process integrates Reinforcement Learning of Human Feedback (RLHF) [252], which enhances the model's ability to follow instructions and minimize the generation of erroneous or harmful outputs. RLHF has applied Proximal Policy Optimization (PPO) [308], a reinforcement learning algorithm demonstrably effective in training an agent's decision-making capabilities to tackle complex tasks.

GPT-4 represents a significant advancement in the GPT series, emerging as a multimodal LLM capable of processing both text and image inputs while generating text outputs. Trained on a massive scale, with the exact number of parameters undisclosed, GPT-4 exhibits human-level performance across different benchmarks, surpassing the capabilities of its predecessor, GPT-3.5.[4] The model's development involved an iterative training process, incorporating adversarial testing and continuous feedback integration. This iterative process has resulted in notable improvements in factual accuracy, steerability (the ability to control and guide the model's output through instructions), and adherence to safety constraints.

Llama is a suite of pre-trained generative text models ranging from 7 billion to 405 billion parameters. We employ both the chat version of the Llama2-7B model and Llama3-70 model optimized for dialogue. The chat

[2]We employ the GPT-3.5-turbo-instruct version.
[3]We employ the NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO version.
[4]The number of parameters in the GPT-4 model is estimated to be about 1.7 trillion.

version prioritizes conversational fluency through further fine-tuning on tens of thousands quick response pairs and additional datasets. To ensure data quality, its training data underwent a rigorous filtering process, which excluded websites with known privacy issues and unreliable sources. Llama training has included RLHF with PPO, enhanced by the strategic assimilation of Rejection Sampling, a technique for acquiring observations from targeted distributions. Llama2 shows little differences with respect to Llama3, with the latter demonstrating notable advancements in computational efficiency and contextual understanding, as well as enhanced multi-turn dialogue consistency. Additionally, Llama3 incorporates a more expansive training dataset and utilizes refined alignment techniques to improve output reliability.

Mistral, an LLM with 7.3 billion parameters, addresses the challenges of processing extensive text and computational efficiency. Sharing a similar size to the previously employed Llama2-7B model, Mistral leverages two key techniques: Sliding Window Attention (SWA) [37, 73] and Grouped Query Attention (GQA) [8]. SWA tackles the computational demands of long sequences by segmenting them into manageable 4096-token windows, effectively reducing memory footprint and accelerating inference. This allows Mistral to handle sequences with a total context length of 32,768 tokens. GQA departs from the conventional attention mechanism by grouping hidden states, leading to significant improvements in efficiency compared to traditional methods. We employ an "instructed" version of Mistral[5], which has been fine-tuned specifically for processing chat-style instructions, incorporating additional optimizations for excelling in conversational interactions.

Mixtral, a composite model combining 8 Mistral-7B models into an

[5]We employ the Mistral-7B-Instruct version.

ensemble (Mixtral-8x7B), represents an innovative approach to leveraging model diversity for improved performance. Fine-tuned with Direct Preference Optimization (DPO) [273], Mixtral achieves notable improvements in both response quality and contextual adaptability. The ensemble model harnesses the strengths of its constituent Mistral models, employing sophisticated alignment and optimization techniques to ensure consistency and coherence in dialogue tasks. By distributing computational loads and combining insights across models, Mixtral not only enhances robustness but also achieves superior reliability in generating nuanced, contextually appropriate outputs. This design underscores its efficacy in managing complex and high-stakes conversational interactions.

To ensure consistency, we adopt the same hyperparameters configuration for all LLMs employed. We set a low-temperature score of 0.10 to promote deterministic outputs, favoring the most probable token selections at each generation step. We set top-p nucleus sampling with a threshold of 0.90. Finally, we establish a maximum output length of 200 tokens and leave all the other parameters with their default values.

To detect moral values from text, we leverage the LLMs' inherent knowledge and contextual comprehension through prompting. As shown in Figure 3.1, we develop two prompting configurations. The first, labeled "single-prompt", assesses the model's capacity for comprehensive in-context reasoning. This configuration presents a single prompt requiring the model to identify whether a sentence conveyed moral content and, if so, to simultaneously identify the specific MFT moral dimensions. This approach tests the model's ability to handle both binary and multi-label moral value detection tasks jointly. The second configuration, labeled "multi-taks", presents a two-step approach with separate instructions to simplify the task. Ini-

tially, the model focuses on determining the presence or absence of moral content in the sentence, answering in a binary format (i.e. "yes" or "no"). If the answer is positive, a subsequent prompt asks the model to tackle the multi-label task, predicting the moral dimensions in the sentence. As shown in Figure 3.1, we design the prompt to encourage the model to generate binary responses (positive or negative) for each moral dimension. This approach aims to reduce the variability in the model's textual output during the detection task.

Building on these configurations, we adapt the structure of the 0-shot single-prompt setting to create the 6-shot and 12-shot few-shot prompts. These few-shot prompts incorporate randomly selected examples from the training set, appended as context to the prompt to guide the model's predictions. The examples are formatted consistently with the single-prompt approach, providing both positive and negative cases for each moral dimension. This design ensures alignment with the 0-shot structure while leveraging in-context learning to improve performance. The random sampling of examples minimizes selection bias and introduces variability, enabling an evaluation of the model's sensitivity to few-shot learning across different moral dimensions. To ensure coverage of all labels, we exclude combinations that do not encompass all labels.

A post-processing step aligns the model's predicted moral dimensions with the MFT framework. Here, the model's identified dimensions are directly mapped to their corresponding MFT labels, except for "Proportionality" and "Equality", which are mapped into "Fairness" to conform with the original version of MFT.

3.1.2 NLI-RoBERTa-based model

Our investigation into unsupervised methodologies for moral value detection explores the potential of NLI systems as zero-shot ready-made classifiers [55]. The NLI model's capacity focuses on determining the relationship between two inputs text, a premise and a hypothesis. This involves determining the degree to which the hypothesis is entailed, contradicts or is neutral w.r.t the premise. Understanding the contextual relationship between text pairs allows NLI models to make inferences about their logical connections. Following [24], we formulate hypotheses for each potential moral value using the input text as the premise. To perform this classification, we leverage pre-trained checkpoints of MNLI-RoBERTa-large [6], a model trained on the MNLI dataset [354].

Our method focuses on two distinct configurations. The first configuration (i.e. 'single-prompt') follows the methodology detailed in [24] (crf. Figure 3.2. Here, we evaluate the entailment score for each value and violation using the pre-trained NLI-RoBERTa model. The entailment score is calculated by normalizing the model's predicted entailment and contradiction probabilities for a given moral value associated to the input text. We consider moral values with a normalized entailment score of 0.50 or higher as strongly associated with the text. Sentences with entailment scores below 0.50 for all moral values are classified as "Non-Moral", signifying a lack of discernible moral content.

The "double-prompt" configuration adopts a two-step approach. In the first stage, we leverage the methodology of [24] to perform a binary classification task. Here, the model assesses whether the input text expresses moral implications by comparing it with the term "moral", treated as a

[6] https://huggingface.co/FacebookAI/roberta-large-mnli

hypothesis. If the entailment score for the hypothesis exceeds a predefined threshold (i.e., 0.50), the text is classified as moral and moves to the second stage. Conversely, the sentence is classified as lacking moral content (i.e. "Non-Moral"), and the process ends. The second stage focuses on identifying moral foundations in text classified as moral in stage one. Following the approach of [55] we normalize entailment and neutrality scores to perform the classification.

3.2 Results and Evaluation

In this section, we present an evaluation of the methods discussed in Section 3.1 in terms of their ability to detect and identify moral content. Specifically, our analysis aims to provide an answer to the following questions:

1. Which model exhibits the superior performance in recognizing moral dimensions within text? How the size of a model or different prompting strategies affect performances?

2. Considering the subjectivity of the task, can we consider the models' performance as satisfactory? How a human would perform on the same task?

Next, we discuss the dataset utilized in our study (i.e. MFRC), with a focus on its pre-processing (Section 3.2.1. Section 3.2.2 presents and discusses a comparative analysis among various models, which aims to elucidate the strengths and weaknesses of each model in the context of moral content identification. Last, in Section 3.2.3, we present an experimental framework designed to facilitate direct comparison between human and model performances.

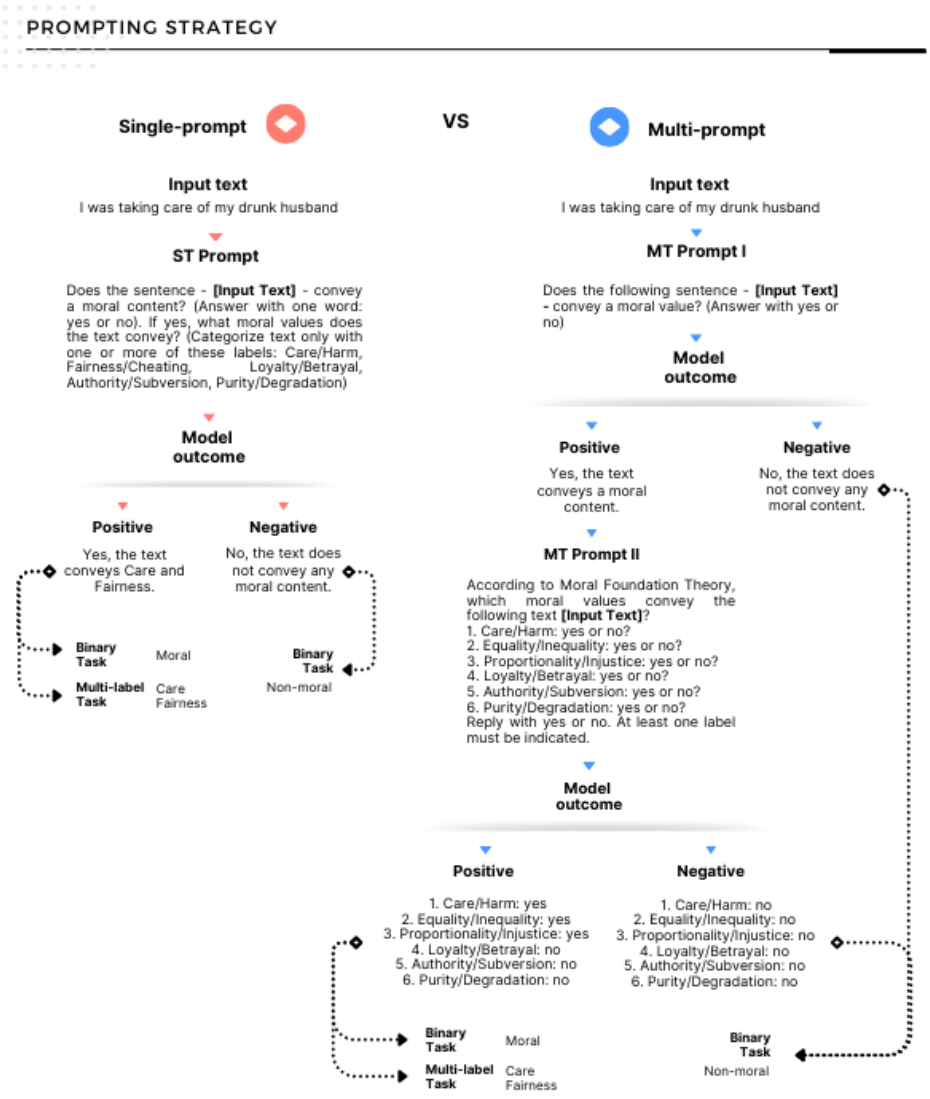


Figure 3.1: Overview of the LLMs-based approach. The single-prompt configuration (left) leverages a single prompt to address both binary and multi-label tasks: identifying the presence of moral content and detecting the specific MFT dimension associated with the text. Conversely, the double-prompt configuration (right) adopts a two-step approach with separate prompts for each task. The first prompt focuses on a binary classification, which identifies whether the text conveys moral content or not. Text classified as moral then progresses to a second prompt that focuses on the detection of one or more MFT dimensions.

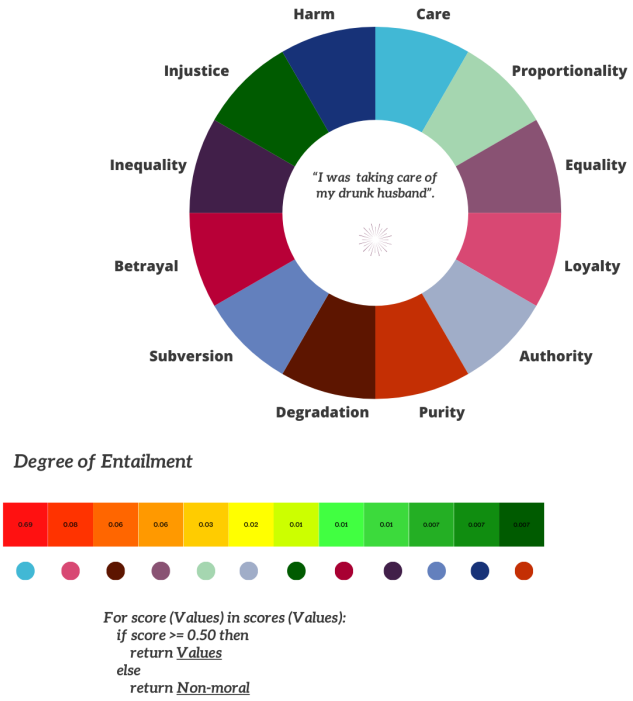


Figure 3.2: Overview of NLI-based method: After processing the input sentence as a premise, the NLI model considers all of the MFT dimensions to be hypotheses, and assesses their connection with each other. As the classification outcomes, we choose labels having an entailment score of at least 0.50. If no label satisfies this requirement, the sentence is classified as lack of moral content ("Non-Moral").

3.2.1 Dataset Preprocessing

To assess the ability of the previously described models in detecting and classifying moral content, we use the MFRC dataset [337], which includes three sub-corpora with 5366, 5351, and 7169 items for the Everyday Moral Life, USA and French politics subcorpus, respectively. Reddit comments sourced from a comprehensive spectrum of 12 distinct subreddit communities. These comments have been annotated across the eight moral dimensions, under the theoretical framework detailed in [27] (crf. Sect. 2.1.1). Diverging from Atari et al. MFT version, we consider the MFT version of [129, 134], in which the dyads of "Equality" and "Proportionality" are grouped into the singular dimension "Fairness". This aggregation allows for a more concise and focused exploration of moral foundations, mitigating redundancy and maintaining the crucial nuances associated with fairness. Furthermore, we exclude the label "Thin Morality", as conceptualized by [28], due to its ambiguous nature. This exclusion arose from the inability to map "Thin Morality" onto any of the five established moral dimensions. Additionally, we take into account comments identified by the majority of annotators as devoid of moral content, designated with the label "Non-Moral". Finally, we exclude entries that have been both labeled as "Non-Moral" and attributed to one or more moral dimensions by annotators, to mitigate potential ambiguity in the corpus. The dataset includes comments from various subreddits, that recall topics about the USA politics, French politics, and everyday moral life. The Everyday Moral Life sub-corpus (i.e. EveryDay) focuses on moral judgments and expressions of moral emotions in non-political contexts, sourced from comments coming from four subreddits. The USA Politics sub-corpus (i.e. USA Politics) comprises comments from three subreddits, encompassing political moral language from diverse perspectives. The French Politics

sub-corpus contains comments from subreddits associated with the French presidential race. Our analysis highlighted a low inter-annotator agreement for this sub-corpus , therefore we exclude it from our assessment. To reduce subjectivity and possible noise, we exclude items classified by a single annotator and by annotators with uncertain confidence levels. Eventually, we obtain $2,837$ items labeled by two to six different annotators for the Everyday Moral Life sub-corpus and $2,757$ items labeled by two to five different annotators for the USA Politics sub-corpus. To establish baseline performance metrics, we aggregated Reddit annotations by majority vote for each moral value, using a threshold of 50%.

3.2.2 Comparison Among Language Models

We evaluate the performance of GPT-3.5, Mistral, Llama2, and NLI-RoBERTa models, on the "Everyday" and "USA Politics" sub-corpora of the MFRC. Our primary focus is to assess their ability to identify moral content and its associated dimensions based on MFT. We conduct a comparative analysis using two different prompting scenarios (crf. Sect. 3.1). In the single-prompt prompt configuration, we discriminate between moral and non-moral content and predict the corresponding dimensions with a single prompt. In the double-prompt prompt configuration, we use a two-stage approach: first, we classify items as moral or non-moral, and then, we associate these items with one or more moral dimensions.

We perform the evaluation from two distinct perspectives. In Section 3.2.2, we explore the models' capability to identify specific moral value dimensions. Conversely, Section 3.2.2 focuses on the ability to detect text that conveys moral content. Although the tools considered address both tasks simultaneously, we discuss the two aspects separately for the sake of clarity. All

experiments have been performed on the same dataset, discussed in Section 3.2.1.

Identification of Moral Value Dimensions

Tables 1 to 6 present the performances of the models in the identification of moral value detection in both single-prompt and double-prompt settings. Each model has been evaluated on the dataset introduced in Section 3.2.1.

Table 3.1 summarizes the overall performance of the models in terms of weighted precision, recall, and F1 score. The metrics have been averaged across the five dimensions of the MFT. Unlike other studies, we do not include a "Non-Moral" dimension to avoid bias. This decision stems from the observation that a significant number of items in the dataset do not convey any moral value and are easier to classify, which could falsely suggest superior performance.

The GPT-4 model demonstrates the highest overall performance, achieving an F1 score of 0.55 across all MFT dyads in the double-prompt configuration. This represents a significant improvement of over 7 percentage points compared to GPT-3.5, which exhibits competitive performance in both single- and double-prompt settings, attaining F1 scores of 0.50 and 0.48, respectively. A prompt analysis reveals that while GPT-4 shows performance comparable to GPT-3.5 in the single-prompt scenario, it demonstrates a substantial performance gain in the double-prompt setting, achieving an increase of over 5 percentage points and surpassing the 0.50 F1 score threshold. The performance decline is due to the binary classification, as discussed in Section 3.2.2. In the single-prompt configuration, GPT-4 exhibits a propensity to classify a greater proportion of content as moral rather than non-moral. This is likely because the model is less sensitive to more com-

plex prompts, potentially interpreting the dyads as constraints that limit the number of positive predictions it can generate.

Llama-70B and Mixtral exihibit F1 scores of 0.45 and 0.43, respectively, under their optimal prompting scenarios. Despite having significantly fewer parameters than GPT-based models and Llama-70B, Mixtral demonstrates competitive performance, achieving F1 scores of 0.43 and 0.41 in the single- and double-prompt settings, respectively. In contrast, Llama-70B achieves slightly higher F1 scores of 0.44 and 0.45 in the single- and double-prompt configurations, respectively.

Notably, Mistral and Llama2 achieve F1 scores of 0.43 and 0.30, respectively, under their optimal prompting configurations. Although Mistral has a significantly lower parameter count compared to Llama-70B, Mixtral, and the GPT-based models, it exhibits competitive performance relative to Llama-7B and is comparable to Mixtral in the single-prompt context. In the double-prompt configuration, Mistral achieves an F1 score of 0.43, although its performance decreases to 0.24 in the single-prompt scenario. These results highlight the importance of prompt complexity in influencing model performance, as models with limited in-context understanding capabilities, as smaller LLMs, may benefit more from the double-prompt scenario. Conversely, Mixtral, Llama-70B and GPT-3.5 demonstrate greater adaptability across both prompt configurations, suggesting its versatility in tackling this task efficiently. GPT-4, characterized by a substantially larger parameter count and enhanced generalization capabilities relative to GPT-3.5, demonstrates superior performance metrics. This improvement is particularly evident in dual-prompt configurations, where it achieves an F1 score that is 5 percentage points higher than GPT-3.5's best result.

Llama-based models exhibit similar performance across both scenarios.

Llama-7B achieves F1 scores of 0.30 and 0.29 in the single- and double-prompt configurations, respectively, while Llama-70B attains F1 scores of 0.44 and 0.45 under the same conditions. In a comparison of models with equivalent parameter configurations, such as Llama-7B and Mistral, both demonstrate analogous performance in the single-prompt configuration. However, in the double-prompt scenario, Mistral exhibits superior performance, surpassing Llama-7B, which showed a notable decline in efficacy under this condition. Furthermore, while Mistral achieves higher precision, indicating its ability to accurately identify relevant moral dimensions, Llama-7B displayed significantly higher recall, suggesting a tendency to identify a broader range of moral dimensions, even potentially irrelevant ones, as shown by a lower precision. This aspect aligns closely with the performance characteristics observed in the Llama-70B model. This trade-off in precision and recall can render Llama-7B less effective, leading to inconsistent results and potentially introducing a positive bias, where the model tends to associate moral values with sentences even for neutral elements. In contrast, Mistral's strength in precision suggests greater selectivity in identifying moral dimensions, resulting in fewer but more accurate outputs, potentially improving the overall quality of predictions.

The NLI-RoBERTa-based model exhibited inferior performance compared to the LLMs, achieving F1 scores of 0.24 and 0.16 in the single- and double-prompt configurations, respectively.

Across the sub-corpora, all models performed significantly better on "EveryDay" content compared to "USA Politics" content. This discrepancy is likely due to the greater difficulty and subjectivity of the tasks in the "USA Politics" sub-corpus, which is supported by the lower level of inter-annotator agreement. Consequently, the subjectivity of these tasks makes

it challenging to establish an objective gold standard, increasing the likelihood of mismatches between the model predictions and the gold standard. This observation is further supported by the analysis of individual annotators, detailed in Section 3.2.3. Among the models, GPT-4 and GPT-3.5 demonstrated the strongest overall performance in MFT detection. GPT-4 achieved F1 scores of 0.61 and 0.49 for the "EveryDay" and "USA Politics" sub-corpora, respectively. Similarly, GPT-3.5 exhibited F1 scores of 0.58 and 0.42 for the "Everyday" and "USA Politics" domains, respectively. Llama-70B, Mixtral, and Mistral exhibit lower, yet comparable performance relative to GPT-based models, considering their reduced parameter configurations. Specifically, under the optimal prompting scenario for the EveryDay subcorpus, these models attain F1 scores of 0.52, 0.51, and 0.51 for Llama-70B, Mixtral, and Mistral, respectively. In the USA Politics subcorpus, the F1 scores are 0.39, 0.36, and 0.34 for Llama-70B, Mixtral, and Mistral, respectively. In contrast, both NLI-RoBERTa and Llama-7B exhibited limitations in this task. This is evident also in the sub-corpora analysis. Specifically, although Llama-7B maintained consistent performance across different prompting scenarios, its F1 scores remained consistently lower than other models.

Tables 3.2 to 3.6 present the detailed performance of all models for every dimension in terms of precision, recall, and F1 score. Notably, GPT-4 and GPT-3.5 consistently demonstrate strong performance, particularly in detecting dimensions like "Care" (i.e. Table 3.2 and "Fairness" (i.e. Table 3.3. Under their optimal prompting scenario, GPT-4 attains F1 scores of 0.69 and 0.53 for these respective dimensions, whereas GPT-3.5 achieves F1 scores of 0.63 and 0.52 for the same dimensions. Furthermore, GPT-based models exhibit a performance difference between the "Everyday" and

"USA Politics" sub-corpora. In the "EveryDay" sub-corpus, which primarily consists of non-politically charged content and benefits from the higher inter-annotator agreement, GPT-4 and GPT-3.5 achieve F1 scores of 0.78 and 0.72 for "Care", respectively. This contrasts with the lower F1 scores of 0.55 and 0.47 obtained in the "USA Politics" sub-corpus. This discrepancy might be attributed to the lower consistency in human judgments in the "USA Politics" sub-corpus compared to the "Everyday" sub-corpus. For the "Fairness" dimension, GPT-based models demonstrate more balanced performance, with GPT-4 and GPT-3.5 achieving F1 scores of 0.52 and 0.54 for the EveryDay subcorpus, and 0.54 and 0.51 for the USA Politics subcorpus, respectively.

Llama-70B, Mixtral and Mistral also show promising results, achieving an overall F1 score of 0.59, 0.55, and 0.55 for the "Care" dimension, with a higher score (i.e. 0.67, 0.66, and 0.65) in the "EveryDay" sub-corpus compared to "USA Politics" (i.e. 0.48, 0.40, and 0.40). The "Fairness" dimension reflects the trend observed in GPT-based models, with Llama-70B and Mixtral exhibiting balanced performance, achieving F1 scores of 0.44 and 0.46 for the EveryDay subcorpus, and 0.43 and 0.46 for the USA Politics subcorpus, respectively. In contrast, Mistral demonstrates superior performance in the EveryDay subcorpus, attaining an F1 score of 0.48, compared to 0.39 in the USA Politics subcorpus. Notably, the double-prompt prompting scenario proves to be beneficial for Llama-70B, Mixtral and Mistral in most cases.

Identifying moral values like "Purity", "Loyalty", and "Authority" presented a greater challenge for all models compared to other dimensions (crf. Tables 3.4- 3.6). GPT-4 achieves F1 scores of 0.37, 0.42, and 0.30 for these dimensions, respectively, under its optimal prompting scenario, while GPT-

Table 3.1: Models overall performance in multi-label moral value detection under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.18	0.24	0.19	0.21	0.18	0.21
	Recall	0.44	0.17	0.44	0.11	0.44	0.14
	F1	0.24	0.20	0.23	0.13	0.24	0.16
Llama-2 (7B)	Precision	0.23	0.25	0.18	0.17	0.20	0.20
	Recall	0.68	0.63	0.65	0.48	0.66	0.55
	F1	0.34	0.35	0.28	0.24	0.30	0.29
Mistral (7B)	Precision	0.31	0.49	0.21	0.35	**0.56**	0.42
	Recall	0.43	0.54	0.30	0.35	0.17	0.45
	F1	0.32	0.51	0.19	0.34	0.24	0.43
Mixtral (8x7)	Precision	0.42	0.36	0.27	0.24	0.34	0.28
	Recall	0.73	0.83	0.63	0.74	0.68	0.78
	F1	0.51	0.49	0.24	0.36	0.43	0.41
Llama-3 (70B)	Precision	0.37	0.39	0.26	0.26	0.32	0.32
	Recall	**0.86**	0.85	0.76	**0.80**	0.81	**0.83**
	F1	0.51	0.52	0.38	0.39	0.44	0.45
GPT-3.5 (175B)	Precision	**0.57**	0.54	0.41	0.38	0.49	0.42
	Recall	0.63	0.63	0.48	0.47	0.55	0.55
	F1	0.58	0.51	0.42	0.42	0.50	0.48
GPT-4 (Unknown)	Precision	0.46	0.55	0.33	**0.42**	0.39	0.48
	Recall	0.81	0.77	0.73	0.64	0.77	0.70
	F1	0.58	**0.61**	0.42	**0.49**	0.50	**0.55**

3.5's scores are 0.28, 0.33, and 0.29, respectively. Further analysis of GPT-4, the best-performing model, revealed that for "Purity" and "Loyalty", the model achieved its strongest performance on the "Everyday" sub-corpus. While, for "Authority" moral dimension, the model achieved its strongest performance on the "USA Politics" sub-corpus. This suggests that references to moral foundations like "Authority" might be more explicitly expressed and frequently encountered in political contexts.

To evaluate the impact of few-shot prompting configurations on the task of moral values detection, we assess the performance of competitive open models of different dimensions in the zero-shot configuration (cf. Table 3.1, specifically Mistral-7B and Mixtral-7x8B. The analysis involves three prompting configurations: zero-shot (no examples in the prompt), 6-shot (six examples from the training set), and 12-shot (twelve examples from the training set). We measure performance variations for each model across individual moral dimensions in terms of F1 score (cf. Figure 3.3. The analysis reveals distinct trends in the performance of Mistral and Mixtral across

Table 3.2: Models performance for predicting the MFT "Care" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.21	0.33	0.17	0.22	0.19	0.21
	Recall	0.55	0.22	0.64	0.18	0.58	0.10
	F1	0.31	0.27	0.27	0.20	0.29	0.13
Llama-2 (7B)	Precision	0.34	0.37	0.22	0.20	0.28	0.30
	Recall	0.78	0.73	**0.78**	0.52	0.78	0.64
	F1	0.47	0.49	0.34	0.29	0.41	0.41
Mistral (7B)	Precision	0.48	0.63	0.35	0.36	0.75	0.52
	Recall	0.51	0.67	0.26	0.43	0.17	0.58
	F1	0.49	0.65	0.30	0.40	0.28	0.55
Mixtral(8x7B)	Precision	0.58	0.49	0.33	0.26	0.47	0.37
	Recall	0.75	0.89	0.51	0.82	0.66	0.86
	F1	0.66	0.63	0.40	0.40	0.55	0.52
Llama-3 (70B)	Precision	0.51	0.53	0.36	0.35	0.45	0.45
	Recall	0.91	**0.92**	0.67	0.77	0.82	**0.86**
	F1	0.65	0.67	0.47	0.48	0.58	0.59
GPT-3.5 (175B)	Precision	**0.76**	0.66	0.55	0.40	**0.69**	0.54
	Recall	0.69	0.72	0.38	0.56	0.57	0.66
	F1	0.72	0.69	0.45	0.47	0.63	0.60
GPT-4 (Unknown)	Precision	0.62	**0.76**	0.45	**0.56**	0.55	**0.69**
	Recall	**0.92**	0.80	0.61	0.54	0.80	0.70
	F1	0.74	**0.78**	0.51	**0.55**	0.65	**0.69**

Table 3.3: Models performance for predicting the MFT "Fairness" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.18	0.19	0.25	0.25	0.21	0.21
	Recall	0.33	0.13	0.31	0.08	0.32	0.14
	F1	0.23	0.15	0.27	0.12	0.25	0.17
Llama-2 (7B)	Precision	0.17	0.18	0.22	0.20	0.20	0.19
	Recall	0.66	0.69	0.72	0.57	0.70	0.62
	F1	0.27	0.28	0.34	0.30	0.31	0.29
Mistral (7B)	Precision	0.14	0.48	0.15	0.45	0.54	0.46
	Recall	0.07	0.48	0.06	0.34	0.16	0.40
	F1	0.09	0.48	0.09	0.39	0.25	0.43
Mixtral (8x7B)	Precision	0.34	0.26	0.34	0.29	0.34	0.27
	Recall	0.71	0.87	0.72	0.86	0.71	0.87
	F1	0.46	0.40	0.46	0.43	0.46	0.41
Llama-3 (70B)	Precision	0.28	0.30	0.25	0.28	0.26	0.29
	Recall	**0.90**	0.81	**0.96**	0.86	**0.93**	0.84
	F1	0.43	0.44	0.39	0.43	0.41	0.43
GPT-3.5 (175B)	Precision	**0.47**	0.38	0.44	**0.46**	**0.45**	0.42
	Recall	0.63	0.63	0.60	0.48	0.62	0.54
	F1	**0.54**	0.47	0.51	0.47	0.52	0.47
GPT-4 (Unknown)	Precision	0.33	0.38	0.31	0.43	0.32	0.40
	Recall	0.86	0.81	0.90	0.76	0.88	0.78
	F1	0.48	0.52	0.46	**0.54**	0.47	**0.53**

Table 3.4: Models performance for predicting the MFT "Purity" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

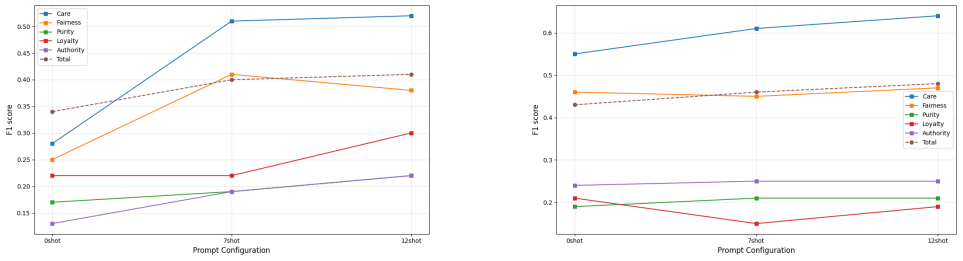
Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.09	0.09	0.09	0.08	0.09	0.09
	Recall	0.50	0.13	0.53	0.07	0.51	0.10
	F1	0.15	0.11	0.15	0.08	0.15	0.09
Llama-2 (7B)	Precision	0.06	0.07	0.07	0.07	0.06	0.07
	Recall	0.16	0.11	0.17	0.08	0.16	0.09
	F1	0.08	0.09	0.10	0.07	0.09	0.08
Mistral (7B)	Precision	0.13	0.18	0.12	0.14	0.41	0.16
	Recall	0.60	0.35	0.46	0.27	0.11	0.31
	F1	0.21	0.24	0.19	0.18	0.17	0.21
Mixtral (8x7B)	Precision	0.14	0.23	0.09	0.16	0.11	0.20
	Recall	**0.77**	0.62	0.58	0.34	0.68	0.49
	F1	0.24	0.33	0.15	0.22	0.19	0.28
Llama-3 (70B)	Precision	0.21	0.19	0.18	0.18	0.20	0.18
	Recall	0.74	0.75	**0.66**	0.59	**0.70**	0.66
	F1	0.33	0.30	0.28	0.27	0.31	0.29
GPT-3.5 (175B)	Precision	**0.29**	0.28	0.23	0.24	0.27	0.26
	Recall	0.35	0.38	0.22	0.21	0.29	0.30
	F1	0.32	0.32	0.22	0.22	0.28	0.28
GPT-4 (Unknown)	Precision	0.28	0.29	**0.32**	0.31	**0.30**	**0.30**
	Recall	0.27	0.58	0.18	0.38	0.23	0.49
	F1	0.27	**0.39**	0.23	**0.34**	0.26	**0.37**

Table 3.5: Models performance for predicting the MFT "Loyalty" dimension in terms of precision. recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model. while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.21	0.15	0.30	**0.32**	0.26	0.20
	Recall	0.08	0.09	0.12	0.06	0.10	0.07
	F1	0.11	0.11	0.17	0.11	0.14	0.11
Llama-2 (7B)	Precision	0.05	0.05	0.05	0.06	0.05	0.05
	Recall	0.90	0.72	**0.83**	0.59	**0.86**	0.65
	F1	0.10	0.09	0.10	0.10	0.10	0.10
Mistral (7B)	Precision	0.11	0.19	0.11	0.18	0.17	0.18
	Recall	0.79	0.46	0.67	0.30	0.32	0.38
	F1	0.19	0.27	0.19	0.22	0.22	0.25
Mixtral (8x7B)	Precision	0.13	0.13	0.12	0.13	0.13	0.13
	Recall	0.79	0.73	0.59	0.54	0.69	0.63
	F1	0.22	0.23	0.20	0.21	0.21	0.22
Llama-3 (70B)	Precision	0.16	0.15	0.16	0.13	0.16	0.14
	Recall	0.82	**0.84**	0.60	0.76	0.71	0.80
	F1	0.27	0.26	0.26	0.22	0.26	0.24
GPT-3.5 (175B)	Precision	0.16	0.23	0.18	0.28	0.17	0.26
	Recall	0.67	0.53	0.62	0.44	0.64	0.49
	F1	0.26	0.32	0.28	0.34	0.27	0.33
GPT-4 (Unknown)	Precision	**0.34**	0.24	0.28	0.28	**0.30**	0.26
	Recall	0.64	0.73	0.73	0.65	0.69	0.69
	F1	**0.44**	0.36	**0.40**	0.39	**0.42**	0.38

Table 3.6: Models performance for predicting the MFT "Authority" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.03	0.03	0.08	0.07	0.05	0.05
	Recall	0.47	0.14	0.48	0.09	0.47	0.10
	F1	0.06	0.05	0.13	0.08	0.10	0.07
Llama (7B)	Precision	0.05	0.06	0.14	0.11	0.10	0.10
	Recall	0.25	0.18	0.28	0.28	0.27	0.26
	F1	0.08	0.09	0.18	0.16	0.14	0.14
Mistral (7B)	Precision	0.07	0.12	0.13	0.26	0.30	0.22
	Recall	0.70	0.14	0.73	0.24	0.09	0.21
	F1	0.12	0.13	0.22	0.25	0.13	0.21
Mixtral (8x7B)	Precision	0.11	0.09	0.16	0.16	0.14	0.14
	Recall	0.55	0.51	0.70	0.55	0.66	0.54
	F1	0.18	0.16	0.26	0.25	0.24	0.22
Llama-3 (70B)	Precision	0.12	0.10	0.15	0.13	0.14	0.12
	Recall	0.42	0.64	0.58	0.85	0.54	0.79
	F1	0.18	0.18	0.24	0.22	0.22	0.21
GPT-3.5 (175B)	Precision	**0.16**	0.12	0.25	**0.26**	**0.22**	0.20
	Recall	0.32	0.36	0.43	0.41	0.40	0.39
	F1	0.21	0.18	**0.32**	**0.32**	0.29	0.27
GPT-4 (Unknown)	Precision	0.10	**0.16**	0.14	0.20	0.13	0.19
	Recall	**0.68**	**0.68**	**0.87**	0.67	**0.82**	0.67
	F1	0.18	**0.26**	0.23	0.31	0.22	**0.30**



(a) Mistral (7B) (b) Mixtral (7x8B)

Figure 3.3: Performance comparison of Mistral-7B and Mixtral-3x8B models in few-shot prompting configurations for moral values detection. The analysis evaluates the impact of zero-shot, 7-shot, and 12-shot configurations on the models' performance across individual moral dimensions in terms of F1 score.

the three prompting configurations, with Mixtral generally outperforming Mistral across most dimensions and configurations. Considering overall performance (i.e. "Total"), Mistral improves from an F1 score of 0.34 in the 0-shot configuration to 0.41 in the 12-shot configuration. Mixtral, however, demonstrates a higher starting point, achieving an F1 score of 0.43 in zero-shot and improving to 0.48 in the 12-shot configuration, underscoring its robustness and stronger overall performance. For the "Fairness" dimension, Mistral exhibits notable improvement, increasing its F1 score from 0.25 in the 0-shot configuration to 0.41 in the 6-shot configuration and 0.38 in the 12-shot configuration. Mixtral, by contrast, starts at a substantially higher F1 score of 0.46 in the 0-shot configuration and maintains stable performance across all configurations, achieving F1 scores between 0.45 and 0.47. In the "Care" dimension, both models benefit significantly from few-shot prompting. Mistral improves from an F1 score of 0.28 in 0-shot to 0.51 in 6-shot and 0.52 in 12-shot. However, Mixtral demonstrates superior performance, starting with an F1 score of 0.55 in zero-shot and further improving to 0.61 and 0.64 in the 6-shot and 12-shot configurations, respectively. The more complex dimensions, such as "Purity" and "Authority", pose challenges for both models. In "Purity", Mistral records F1 scores ranging from 0.17 in 0-shot to 0.22 in 12-shot, while Mixtral performs slightly better, achieving F1 scores between 0.19 and 0.21 across configurations. For "Authority", Mistral starts at 0.13 in 0-shot and improves to 0.22 in 12-shot, whereas Mixtral consistently outperforms Mistral, with F1 scores between 0.24 and 0.25. The "Loyalty" dimension follows a similar pattern. Mistral shows progressive improvement from 0.22 in 0-shot to 0.30 in 12-shot. In contrast, Mixtral starts at 0.21 in 0-shot but exhibits a slight decline in the few-shot configurations, with scores of 0.15 and 0.19 in 6-shot and 12-shot, respectively.

Overall, these results demonstrate that Mixtral generally achieves better performance than Mistral, particularly in the more interpretable dimensions such as "Fairness" and "Care". Both models, however, benefit from few-shot prompting configurations, with improvements observed in the more challenging dimensions compared to the 0-shot baseline. Notably, Mistral exhibits a more substantial relative improvement across configurations, with a 7-point gain in the 12-shot configuration compared to its 0-shot baseline, compared to a 5-point gain for Mixtral. This highlights the responsiveness of smaller models, such as Mistral, to few-shot learning strategies.

Detection of Text with Moral Content

We assess the effectiveness of all models in discerning whether a text conveys or not moral content (as a binary variable). We evaluate both LLMs and the NLI-RoBERTa model performance for both the "Everyday" and "USA Politics" sub-corpora of the MFRC. Items annotated with at least one moral dimension by the majority of human annotators are considered positive cases (they convey moral content), while the remaining are considered as negative cases (lacking moral content).

Table 3.7 summarizes the overall performance in terms of accuracy, weighted precision, recall, and F1 score.

Among all models, GPT-4 achieves the highest performance under its optimal prompting configuration. It exhibits F1 and accuracy scores of 0.81 and 0.86, respectively. GPT-3.5 follows closely, achieving F1 and accuracy scores of 0.76 and 0.83, respectively. Llama2 achieves the lowest score among the LLMs, with F1 and accuracy scores of 0.52 for both metrics. Excepting GPT-3.5, all models exhibit a performance increase with the double-prompt prompting configuration. This configuration significantly improves Llama-

70B, Mixtral and Mistral's performance, resulting in an accuracy of 0.67, 0.74 and 0.77 and an F1 score of 0.76, 0.67 and 0.62. The NLI-RoBERTa-based model exhibits the lowest overall performance among the evaluated models, achieving F1 and accuracy scores of 0.45 and 0.62, respectively, even under its optimal prompting configuration. Interestingly, the model demonstrates a slight improvement in moral content prediction accuracy in the double-prompt setting compared to the single-prompt configuration. This discrepancy might be attributed to the presence of non-moral content impacting the single-prompt metric calculations, potentially suggesting the model's enhanced ability to identify moral content under the double-prompt framework.

Subcorpus-level analysis reveals a significant performance disparity across all models, with superior results consistently observed in the "Everyday" subcorpus compared to the "USA Politics" subcorpus. GPT-4 demonstrates robust performance across both subcorpora. Under optimal prompting conditions, it achieves accuracy scores of 0.90 and 0.81, alongside F1 scores of 0.87 and 0.74, for the "Everyday" and "USA Politics" subcorpora, respectively. Among the smaller models, GPT-3.5 outperforms Llama-70B by 10 percentage points in accuracy, and by 6 percentage points when compared to Mistral and Mixtral. In terms of F1 score, GPT-3.5 leads by 8, 6, and 12 points for the "Everyday" subcorpus. For the "USA Politics" subcorpus, GPT-3.5 maintains this trend, achieving an accuracy of 0.80, compared to 0.56, 0.67, and 0.73 for Llama-70B, Mixtral, and Mistral, respectively, and an F1 score of 0.70, compared to 0.59, 0.65, and 0.53 for the same models.

Among the less parameterized models, Mistral consistently outperforms Llama-7B. In the "Everyday" corpus, the F1 score and accuracy differences are 16% and 25% points, respectively, while in the "USA Politics" subcorpus,

Table 3.7: Models overall performance in binary moral value detection under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.33	0.47	0.35	0.42	0.34	0.45
	Recall	0.64	0.25	0.71	0.18	0.68	0.21
	F1	0.43	0.32	0.47	0.25	0.45	0.29
	Accuracy	0.38	0.63	0.43	0.62	0.41	0.62
Llama (7B)	Precision	0.37	0.44	0.35	0.37	0.36	0.41
	Recall	**0.98**	0.72	0.96	0.60	**0.97**	0.66
	F1	0.54	0.55	0.51	0.46	0.52	0.50
	Accuracy	0.38	0.56	0.36	0.48	0.37	0.52
Mistral (7B)	Precision	0.86	0.82	0.67	0.68	0.78	0.76
	Recall	0.46	0.62	0.24	0.43	0.35	0.53
	F1	0.60	0.71	0.35	0.53	0.49	0.62
	Accuracy	0.78	0.81	0.69	0.73	0.73	0.77
Mixtral (8x7B)	Precision	0.93	0.90	0.90	0.87	0.92	0.89
	Recall	0.66	0.68	0.48	0.52	0.56	0.59
	F1	0.77	0.77	0.63	0.65	0.70	0.71
	Accuracy	0.81	0.80	0.67	0.63	0.74	0.71
Llama (70B)	Precision	**0.99**	0.95	**0.99**	0.91	**0.99**	0.93
	Recall	0.53	0.62	0.39	0.44	0.45	0.52
	F1	0.69	0.75	0.56	0.59	0.62	0.67
	Accuracy	0.68	0.77	0.45	0.56	0.57	0.67
GPT-3.5 (175B)	Precision	0.82	0.90	0.65	0.79	0.73	0.85
	Recall	0.84	0.70	0.76	0.57	0.80	0.64
	F1	0.83	0.79	0.70	0.66	0.76	0.73
	Accuracy	0.87	0.86	0.77	0.80	0.82	0.83
GPT-4 (Unknown)	Precision	**0.99**	0.88	**0.99**	0.78	**0.99**	0.83
	Recall	0.69	0.87	0.48	0.71	0.57	0.79
	F1	0.81	**0.87**	0.65	**0.74**	0.72	**0.81**
	Accuracy	0.83	**0.90**	0.62	**0.81**	0.73	**0.86**

the differences are 2% and 25% points for the F1 score and accuracy, respectively. In general, the overall disparity between F1 score and accuracy for each model is noteworthy. This is due to the imbalance between moral and non-moral items. F1 score summarizes precision and recall in predicting positive (moral) items, whereas accuracy considers both moral and non-moral items (the latter more prevalent in the dataset). The NLI-RoBERTa-based model exhibits consistent performance across both subcorpora; however, its scores remain significantly lower compared to the other models.

3.2.3 Human vs. Models Comparison

We present a framework for the evaluation of machine learning models in tasks characterized by high subjectivity and low inter-annotator agreement. This framework fosters a direct comparison between human and model per-

formance by juxtaposing annotators' responses with model predictions. In short, we exclude an annotator from the calculation of the gold-standard annotations and consider them as a human predictor to be compared with model performance. In this way we ensure that the "human predictor" does not influence the gold standard. We repeat this process for each annotator and present both individual and aggregated results. The MFRC dataset employs six annotators for the "EveryDay" subcorpus and five for the "USA Politics" subcorpus. Since the "USA Politics" subcorpus provides comparable insights, we exclusively report the performance of the "EveryDay" subcorpus. To address variations in annotator coverage across dataset sentences, we employ six distinct versions of the "EveryDay" subcorpus. Each version focuses on items tagged by a single annotator, ensuring a minimum of three annotators per item. Within each version, each element is associated with two values: a distinct value assigned by the specific annotator (human predictor) and an aggregate value from the remaining annotators. The aggregate value reflects the majority consensus, determined by ratings exceeding or equaling a 50% threshold and it represents the gold standard (or ground truth) for that element. Our results are summarized in three tables: Table 3.8 shows the results of comparing human annotators and models in a multi-label value detection scenario. In Table 3.9, the performance attained by the best annotator for each moral foundation is juxtaposed with the model's performance. Finally, Table 3.10 outlines the results of our comparison in a binary value detection task, focused on identifying the morality in the sentence.

Table 3.8 presents the performance of all models in a multi-label moral value detection task based on the five MFT dimensions. We analyze the results for each dyad in terms of weighted F1 score, which takes into account

annotator support, and average F1 score, which disregards such support. The results indicate that GPT-4 and GPT-3.5 consistently outperform human annotators in three of the five moral dimensions ("Care", "Fairness", and "Purity"). Additionally, GPT-4 demonstrates superior performance in one additional dimension ("Loyalty"), highlighting its broader capability in this context. For the "Care" dyad, GPT-4 achieves the highest weighted F1 score (i.e. 0.83), whereas GPT-3.5 attains the best average F1 score (i.e. 0.77) compared to the human average (i.e. 0.73). [7] These results translate into a performance advantage of 4-10 percentage points. Similarly, GPT-3.5 shows superior performance in the "Fairness" (4-5 percentage point improvement) while GPT-4 outperforms across the "Purity"and "Loyalty" dimensions, achieving 9–10 and 5 percentage point improvements, respectively, as measured by F1 scores. Conversely, human annotators exhibit superior performance compared to GPT-based models in the "Authority" dimension regarding average F1 scores, with humans scoring 0.22 compared to GPT-4's 0.18. However, in terms of weighted F1, GPT-4 surpasses human performance (i.e. 0.25 vs. 0.20). In general, the results suggest that GPT-3.5 may be more proficient in the single-prompt prompting configurations, where it achieves comparable performance to human judgment in classifying individual value dyads. GPT-4, however, achieves superior performance across both single- and double-prompt configurations, matching human performance on three out of five dyads and yielding significant improvements in the double-prompt setting for the remaining two. Notably, GPT-3.5 excels in well-defined domains such as "Care" and "Fairness", demonstrating statistically significant improvements over human evaluation. However, the model has a drop in performance in more ambiguous domains, mirroring

[7] we discuss this counter-intuitive result in Section 3.3

the complexities faced by human judgment in these areas. This is further supported by F1 scores falling below 0.50 for these MFT dimensions. Conversely, GPT-4 demonstrates consistent robustness, surpassing both GPT-3.5 and human annotators in these less well-defined domains, particularly for "Purity", "Loyalty", and "Authority". Among smaller models, Llama-70B exhibits competitive performance relative to Mixtral in the double-prompt configuration, with weighted average F1 scores of 0.57 and 0.54, respectively, and an identical average F1 score of 0.48. Despite this, both models underperform relative to human evaluators, with Llama-70B trailing human performance by a margin of 2–5 percentage points in weighted F1 scores. Similarly, Mistral achieved proficiency in the double-prompt prompting configuration, obtaining a weighted average F1 score of 0.54 and an average F1 score of 0.47. However, its performance falls short of human evaluation by a margin of 5 percentage points in the overall weighted F1 score. Conversely, Llama2 and NLI-RoBERTa models underperform, with weighted average and average F1 scores of 0.38 and 0.27, respectively, compared to the human benchmark of 0.59.

Table 3.9 provides a comparison between the top-performing human annotators and all models in terms of precision, recall, and F1 scores. Each MFT dimension is assessed in comparison with the best-performing human annotator. GPT-4 achieves an F1 score within 2 percentage points of the best human, indicating its proficiency in this task. GPT-3.5 follows closely, achieving an F1 score within 5 percentage points. In contrast, Llama-70B exhibits a 13 percentage point gap, while Mixtral and Mistral lag by 16 and 15 percentage points, respectively. Llama2 and NLI-RoBERTa show significantly larger gaps, trailing by 33 and 44 percentage points, respectively. These results are based on the optimal prompting configuration for

each model. Specifically, Table 3.9 reveals notable variations in performance across MFT dimensions. In dyads considered to be more readable, such as "Care" and "Fairness", the performance of the best model (i.e. GPT-4) more closely reflects human evaluation. The model achieves F1 scores of 0.87 and 0.62 on these moral dimensions, compared to human averages of 0.87 and 0.61. In contrast, GPT-4 achieves F1 scores of 0.39, 0.25, and 0.29 in "Loyalty", "Authority", and "Purity", respectively. These scores are lower than human performance in these categories, which average 0.49, 0.67, and 0.40, respectively. GPT-3.5 shows a marginally better performance than GPT-4 on "Authority"and "Purity", with F1 scores higher by 2–4 percentage points, respectively. However, it underperforms relative to GPT-4 across the remaining dimensions. When comparing smaller models, Llama-70B outperforms Mixtral by 3–4 percentage points in "Care", "Fairness", and "Loyalty". Conversely, Mixtral demonstrates superior performance in "Authority"and "Purity", with gains of 4–9 percentage points over Llama-70B. Notably, Mistral shows comparable performance to Mixtral in the "Care"and "Authority"dyads, with respective gaps of 2% and 15%. The double-prompt configuration enhances Mistral's performance, enabling it to outperform in certain contexts. Additionally, we revisit a previously observed trend (crf. Sect. 3.2.2 and Sect. 3.2.2 in model behavior, which reflects their inherent characteristics. Llama2 exhibits a bias towards high recall but lower precision, while Mistral demonstrates a tendency to predict fewer labels with greater precision. Notably, Mistral achieves a remarkably high overall precision score (i.e. 0.54 in the double-prompt configuration). However, the double-prompt prompting configuration in Mistral fails to yield optimal results, as it predicts fewer labels despite maintaining high precision.

Table 3.10 shows the results of the binary moral detection task (crf.

Sect. 3.2.2 in terms of moral F1 score and accuracy. Notably, GPT-4 exhibits high performance, achieving scores closer to human annotators across most cases for both F1 score and accuracy. GPT-4 obtained a noteworthy overall F1 score of 0.81 and an overall accuracy of 0.94, closely following human performance of 0.84 F1 score and 0.88 accuracy. Similarly, GPT-3.5 demonstrates competitive performance with F1 and accuracy scores of 0.79 and 0.91, respectively. It is worth noting that the best-performing human annotator (i.e. "Annotator01") achieves F1 and accuracy scores of 0.92 and 0.95, respectively. This outperforms the best-performing model (i.e. GPT-4 in its best-prompting scenario) by a narrow margin of 2 and 1 percentage points, respectively. Finally, the analysis of individual annotators reveals a specific pattern associated with "Annotator00". This annotator predominantly assigns labels indicating the absence of moral content and leaves most of the items with moral content without annotations. Therefore the associated dataset is unbalanced towards items that do not convey moral content. As a result the accuracy is very close to 1 (approximated to 1 in Table 3.10 while the F1-score is significantly lower (0.67) since this metric is influenced by the low recall.

Table 3.8: Direct comparison of human and model performance in multi-label value detection task based on the MFT dimensions. Performance is evaluated using both weighted F1 score (i.e. considering annotator support) and average F1 score (i.e. disregarding annotator support). Performance refers to the MFRC's EveryDay subcorpus.

Dyad	Metrics	Annotator	NLI (355M)		Llama (7B)		Mistral (7B)		Mixtral Nous Hermes (8x7B)		Llama (70B)		GPT-3.5 (175B)		GPT-4(Unknown)	
			Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Care	Weighted avr.	0.73	0.31	0.30	0.47	0.53	0.39	0.71	0.72	0.68	0.73	0.73	0.78	0.73	0.80	**0.83**
	Avg.	0.73	0.27	0.26	0.41	0.46	0.39	0.64	0.62	0.58	0.62	0.62	**0.77**	0.67	0.68	0.75
Fairness	Weighted avr.	0.53	0.26	0.18	0.26	0.30	0.28	0.45	0.48	0.46	0.47	0.49	**0.58**	0.46	0.54	0.54
	Avg.	0.44	0.23	0.15	0.21	0.24	0.25	0.37	0.39	0.36	0.38	0.39	**0.48**	0.38	0.43	0.44
Loyalty	Weighted avr.	0.35	0.27	0.15	0.12	0.10	0.16	0.20	0.26	0.28	0.30	0.28	0.31	0.30	**0.40**	0.39
	Avg.	0.29	0.22	0.11	0.10	0.09	0.14	0.17	0.21	0.22	0.25	0.23	0.26	0.25	**0.34**	0.33
Authority	Weighted avr.	0.21	0.06	0.08	0.05	0.13	0.20	0.16	0.25	0.22	0.21	0.23	0.19	0.18	**0.25**	0.23
	Avg.	**0.22**	0.05	0.06	0.04	0.08	0.18	0.16	0.17	0.12	0.13	0.15	0.16	0.14	0.17	0.18
Purity	Weighted avr.	0.29	0.13	0.07	0.06	0.05	0.13	0.16	0.21	0.28	0.29	0.26	0.24	0.35	0.13	**0.39**
	Avg.	0.24	0.11	0.06	0.05	0.05	0.12	0.13	0.17	0.23	0.25	0.21	0.22	0.29	0.12	**0.33**
Total	Weighted avr.	0.59	0.27	0.23	0.34	0.38	0.31	0.54	0.49	0.54	0.49	0.57	0.63	0.57	0.56	**0.66**
	Avg.	0.57	0.24	0.20	0.29	0.33	0.31	0.47	0.48	0.45	0.48	0.48	**0.60**	0.52	0.54	**0.60**

Table 3.9: Direct comparison between the top performing human annotator and models' performance for each MFTs dimensions. Performance is evaluated in terms of precision. recall and F1 score. Performance refers to the MFRC's EveryDay subcorpus.

Dyad	Metrics	Annotator	Nli (355M)		Llama (7B)		Mistral (7B)		Mixtral Nous Hermes (8x7B)		Llama (70B)		GPT-3.5 (175B)		GPT-4(Unknown)	
			Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Care	Precision	0.86	0.23	0.40	0.34	0.41	**0.89**	0.73	0.64	0.53	0.62	0.61	0.88	0.71	0.73	0.86
	Recall	0.88	0.63	0.31	0.74	0.76	0.28	0.76	0.84	0.92	**0.98**	0.97	0.80	0.80	0.98	0.88
	F1	**0.87**	0.33	0.35	0.46	0.53	0.42	0.75	0.73	0.67	0.76	0.75	0.83	0.75	0.84	**0.87**
Fairness	Precision	0.85	0.22	0.28	0.19	0.23	**0.93**	0.53	0.40	0.33	0.35	0.39	0.56	0.47	0.44	0.49
	Recall	0.48	0.33	0.15	0.59	**0.73**	0.17	0.47	0.65	**0.89**	0.84	0.84	0.62	0.70	0.84	0.84
	F1	0.61	0.27	0.19	0.29	0.35	0.29	0.50	0.50	0.49	0.49	0.53	0.59	0.56	0.58	**0.62**
Loyalty	Precision	**0.47**	0.21	0.38	0.06	0.04	0.15	0.11	0.14	0.14	0.17	0.16	0.22	0.22	0.32	0.27
	Recall	0.59	0.39	0.17	**0.94**	0.56	0.33	0.28	0.83	0.78	0.78	0.83	0.83	0.56	0.50	0.72
	F1	**0.49**	0.27	0.23	0.11	0.08	0.20	0.16	0.23	0.24	0.28	0.27	0.35	0.32	0.39	0.39
Authority	Precision	**0.50**	0.04	0.05	0	0.10	**0.50**	0.29	0.13	0.05	0.10	0.11	0.18	0.12	0.15	0.14
	Recall	1	0.50	0.25	0	0.25	0.25	0.50	0.50	0.25	0.25	0.50	0.50	0.50	0.75	0.75
	F1	**0.67**	0.07	0.08	0	0.14	0.33	0.36	0.21	0.09	0.14	0.17	0.27	0.19	0.25	0.23
Purity	Precision	**0.45**	0.06	0.10	0	0	0.33	0.12	0.31	0.17	0.11	0.10	0.14	0.25	0.14	0.21
	Recall	0.36	0.43	0.14	0	0	0.07	0.29	**0.86**	0.64	0.50	0.50	0.14	0.50	0.14	0.50
	F1	**0.40**	0.11	0.12	0	0	0.12	0.17	0.18	0.27	0.18	0.16	0.14	0.33	0.14	0.29
Total	Precision	**0.73**	0.19	0.32	0.23	0.28	0.69	0.54	0.46	0.39	0.45	0.45	0.66	0.52	0.55	0.63
	Recall	0.71	0.59	0.25	0.66	0.66	0.23	0.60	0.79	0.88	**0.90**	0.89	0.71	0.72	0.85	0.84
	F1	**0.71**	0.27	0.27	0.33	0.38	0.32	0.56	0.55	0.52	0.58	0.58	0.66	0.59	0.65	0.69

Table 3.10: Direct comparison of human and model performance in the binary moral detection task in terms of accuracy, precision, recall, and F1 score. Performance refers to the MFRC's EveryDay subcorpus.

Annotators	Metrics	Annotator	Nli (355M)		Llama (7B)		Mistral (7B)		Mixtral Nous Hermes (8x7B)		Llama (70B)		GPT-3.5 (175B)		GPT-4(Unknown)	
			Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Annotator00	F1	**0.67**	0.02	0.04	0.01	0.03	0.29	0.18	0.04	0.05	0.02	0.04	0.29	0.40	0.06	0.20
	Accuracy	**1**	0.18	0.84	0.04	0.56	0.98	0.97	0.85	0.87	0.69	0.83	0.97	0.99	0.91	0.97
Annotator01	F1	**0.92**	0.40	0.38	0.49	0.55	0.64	0.78	0.77	0.78	0.69	0.78	0.87	0.83	0.84	0.90
	Accuracy	**0.95**	0.36	0.68	0.35	0.61	0.82	0.87	0.82	0.83	0.72	0.82	0.92	0.90	0.89	0.94
Annotator02	F1	0.39	0.44	0.37	0.53	0.56	0.64	0.76	0.80	0.81	0.73	0.80	**0.90**	0.85	0.88	0.90
	Accuracy	**0.93**	0.38	0.66	0.37	0.58	0.81	0.85	0.84	0.85	0.74	0.83	**0.93**	0.91	0.90	0.93
Annotator03	F1	0.68	0.47	0.31	0.53	0.58	0.58	0.72	0.76	0.78	0.67	0.77	**0.83**	0.76	0.80	0.90
	Accuracy	0.83	0.40	0.63	0.38	0.60	0.78	0.83	0.83	0.85	0.73	0.83	**0.88**	0.86	0.91	0.94
Annotator04	F1	0.73	0.50	0.38	0.59	0.61	0.66	0.78	0.81	0.80	0.75	0.80	**0.86**	0.84	0.84	0.89
	Accuracy	0.70	0.44	0.62	0.43	0.60	0.77	0.83	0.81	0.80	0.72	0.79	**0.88**	**0.88**	0.84	0.91
Annotator05	F1	**0.90**	0.49	0.35	0.66	0.67	0.65	0.82	0.88	0.90	0.83	0.86	0.88	0.86	0.89	0.94
	Accuracy	**0.90**	0.42	0.57	0.51	0.64	0.74	0.85	0.87	0.90	0.81	0.85	0.89	0.88	0.88	0.94
Total	F1	0.84	0.40	0.32	0.47	0.51	0.59	0.69	0.70	0.71	0.63	0.70	**0.79**	0.77	0.76	0.81
	Accuracy	0.88	0.37	0.67	0.34	0.60	0.82	0.86	0.83	0.84	0.73	0.82	**0.91**	0.90	0.89	0.94

3.3 Discussion

Our study investigates the performance of LLMs in identifying moral content and its corresponding MFT dimensions in both zero-shot and few-shot settings. We compare GPT-4, GPT-3.5, Mitral, Mistral, Llama-70B, Llama-7B, and NLI-RoBERTa under different prompting configurations (i.e. single-prompt vs. double-prompt) and evaluate them against human annotators. The zero-shot approach ensures generalizability by preventing the models from overfitting to a specific training dataset. Consequently, the evaluation focuses on the LLMs' inherent ability to recognize and correctly contextualize moral concepts. In contrast, the few-shot approach assesses the models' potential to improve performance with minimal guidance.

GPT-4 emerges as the strongest performer, closely followed by GPT-3.5, achieving the highest accuracy in both moral content detection and multi-label classification of MFT dimensions. This likely stems from its larger

size and superior contextual understanding capabilities. These features allow GPT-4 to grasp complex moral concepts without specific training and make subtle distinctions between different MFT dimensions. Interestingly, GPT-4's performance in the double-prompt configuration outperforms that of a single human annotator, on average, suggesting its potential to predict the aggregate human assessment of moral content. GPT-3.5 also outperforms human annotators, albeit to a lesser degree than GPT-4, particularly in the single-prompt configuration. The single-prompt configuration, while computationally less demanding, demonstrates robust performance, especially with smaller models. However, the double-prompt configuration proves more effective with larger models such as GPT-4, leveraging their advanced capabilities to achieve superior results.

Human annotator analysis yields surprising results. GPT-4 performs remarkably similarly to human annotators. The model exhibits internal consistency across all sub-corpora derived from human data, with minimal variability for each category. Furthermore, GPT-based models' overall performance seems to be superior to that of a single human annotator, especially for GPT-4 model. It might seem counter-intuitive that a tool outperforms humans in moral value detection since human values are defined by us. Therefore, we expect the human judgment to be "correct" by definition. However, given the subjectivity of the task, the gold standard is defined by averaging human judgments, therefore the superior performance of GPT-based models might suggest that its prediction resembles the average judgment of humans more closely than the judgment of a single average human. Notably, models excel at predicting high-level value dimensions like "Care" and "Fairness". However, they struggle with more specialized concepts like "Purity", "Loyalty", and "Authority". Interestingly, even humans

face similar difficulties in these areas, suggesting these dimensions may require theoretical reconsideration due to their inherent ambiguity. As shown for the multi-label scenario, GPT-based models remain the top performer even in the binary moral content detection task, achieving results that match or surpass those of humans.

Among the smaller LLMs, Llama-70B demonstrates comparable performance to Mixtral in the single-prompt scenario and slightly surpasses Mixtral in the double-prompt scenario. This superior performance is likely attributable to Llama-70B's larger size relative to Mixtral. Regarding the GPT-3.5 model, the single-prompt configuration appears to be more effective than the double-prompt configuration, with the exception of Llama-70B, which exhibits similar performance across both prompting strategies. This observation highlights that these models can leverage their in-context learning capabilities to perform well even with a single prompt, thereby obviating the need for double interrogation. In contrast, the double-prompt configuration significantly enhances performance for larger models such as GPT-4 and smaller models like Mixtral and Llama-7B, underscoring the varying dependencies on prompting strategies based on model size and architecture.

Mistral outperforms Llama2 in both single-prompt and double-prompt settings, despite having the same size and training conditions. Additionally, the double-prompt configuration benefits Mistral more, potentially mitigating limitations in its in-context understanding capabilities. Mistral's responses are more concise and accurate, achieving high performance compared to humans in both binary and multi-label scenarios. In contrast, Llama2 appears not to be suitable for this task, exhibiting a bias toward predicting many labels and generating overly verbose and inconsistent responses. As expected, smaller models underperform both GPT-4 and GPT-

3.5 across all tasks and prompting configurations. This is likely due to GPT-based models' larger size, granting them access to a broader range of implicit knowledge for understanding moral nuances. However, Mistral emerges as an alternative for binary and multi-label detection of moral values due to its good performance with a smaller size (7B) and lower computational requirements compared to GPT-4 and GPT-3.5 (175B). NLI-RoBERTa consistently underperforms, likely due to its smaller size and training focused on a specific task (entailment).

Subcorpus analysis reveals that all models perform better on the "Everyday" subcorpus compared to the "USA Politics" subcorpus. This is likely due to the higher level of inter-annotator agreement and the more straightforward moral content in the "Everyday" corpus. The "USA Politics" subcorpus, on the other hand, presents greater challenges due to its complexity and the presence of more ambiguous moral concepts.

The evaluation of LLM performance in a few-shot context reveals significant benefits associated with this configuration, particularly for smaller models such as Mistral. Notably, performance improves progressively as the number of examples increases, transitioning from a zero-shot setting to configurations incorporating 6 and 12 examples within a single-prompt scenario. These results underscore the models' ability to effectively utilize in-context learning, with smaller-scale models demonstrating substantial performance gains when provided with a limited set of task-specific examples.

Our findings have several important implications. First, they demonstrate the potential of LLMs for moral assessment, suggesting that they can be used reliably to identify and understand moral content in text. Second, they highlight the importance of model size and in-context understanding capabilities for moral content detection. Third, they suggest that the

double-prompt configuration can be a beneficial prompting approach for smaller LLMs. Finally, they emphasize the need for further research on moral content detection in more complex and challenging contexts, such as the "USA Politics" subcorpus. Additionally, the performance on more ambiguous dimensions (i.e., "Purity", "Loyalty", and "Authority") of both models and humans, highlights these concepts' inherent ambiguity. The low inter-annotator agreement among humans suggests a potential need for a theoretical reevaluation of these dimensions.

4 A Novel Framework for Evaluating Classifiers in a Multi-perspective Domain

The evolution of human-centered AI systems demands reevaluating traditional validation paradigms that treat human annotations and machine learning models as separate entities. While human annotators remain essential for creating benchmark datasets, modern workflows increasingly position AI models as active collaborators in annotation processes [155] (cf. Sect. 1.3.2 and Sect. 2.2.2). This symbiotic relationship necessitates validation frameworks that operate on a unified plane, where human and machine performance can be directly compared through interpretable metrics. Current evaluation methodologies suffer from a fundamental dichotomy: machine learning models are assessed through detection-oriented metrics like F1-score, while human annotation quality is measured through agreement statistics such as Fleiss' Kappa. This creates an artificial separation between two intrinsically connected aspects of AI development. The F1-score optimizes for precision-recall tradeoffs but ignores task subjectivity, while Fleiss' Kappa measures consensus without quantifying concrete detection performance against a ground truth. This disparity impedes meaningful comparisons and obscures opportunities for human-AI quality synergies. A key challenge in model evaluation is the lack of a standardized framework for directly comparing machine learning performance with human annotation

patterns. Traditional inter-annotator agreement (IAA) metrics like Fleiss' Kappa, while valuable for assessing annotation consistency, are not suitable for evaluating model performance. Conversely, model evaluation metrics provide no insights into how machine performance align with human annotation variability. This methodological gap is particularly problematic in subjective annotation tasks where model interpretability gaps and annotation uncertainty hinder robust validation. An example of this evaluative discrepancy emerges in the domain of moral value detection, a task that is inherently affected by subjectivity and interpretive ambiguity. In a recent model assessment conducted on a benchmark dataset in this field [53], the best-performing zero-shot large language model (i.e. GPT-4) achieves an overall F1-score of 0.55. However, the inter-annotator agreement, as measured by Fleiss' Kappa, is 0.36. Since the two values cannot be compared, whether the model's performance is satisfactory is unclear, considering that even human would not perform greatly in the same task. Other significant areas of application include sentiment and emotion analysis, as well as the investigation of social norms and human behavior through machine learning models. These domains are often hindered by a lack of comprehensive benchmark datasets, and are characterized by highly subjective annotations, which consequently result in low inter-annotator agreement.

Presenting the contribution of Bulla et al. [58], we propose a unified validation framework that synthesizes model performance metrics and human agreement measures into a single evaluative plane. Our approach introduces a hybrid metric, F1-kappa, which harmonizes the precision-recall balance of F1-score with Fleiss' Kappa's expectation-adjusted design. By embedding agreement uncertainty into model evaluations (and vice versa), F1-kappa enables direct comparison between human and machine performance while

quantifying their alignment. This metric addresses two critical limitations of current methodologies: the inability to compare model outputs with human annotations using a common scale, and the flexibility to accommodate incomplete annotation distributions and varying levels of annotator participation, thus facilitating a more accurate and comprehensive assessment of dataset heterogeneity.

We validate our approach through large-scale experiments spanning binary, multi-class, and multi-label classification tasks on synthetic and real datasets. Results demonstrate that the F1-kappa behave similarly to Fleiss' Kappa at various conditions. Furthermore F1-kappa supports any kind of settings including multi-label (more than one choice per annotator) and can focus on a specific class (e.g., positive samples), besides making human agreement comparable with model performance.

The Chapter is structured as follows. Section 4.1 briefly introduces Fleiss' Kappa, then describes our F1-kappa metric, while Section 4.2 present results on both synthetic and real data. We present an in-depth state-of-the-art analysis in Chapter 2.2.2.

4.1 A Unified Approach to Inter-Annotator Agreement and Evaluation

The proposed metric adapts the F1 score to measure inter-annotator agreement with more than two annotators while accounting for agreement by chance, taking inspiration from the popular Fleiss' Kappa score. In the following sections, we first describe Fleiss' Kappa, then introduce our F1-kappa score. Finally, we detail how to make the performance of a classifier comparable to F1-kappa.

4.1.1 Background

Perhaps the most known metric for measuring inter-annotator agreement for categorical data is Fleiss' Kappa [110]. Introduced by Joseph L. Fleiss in 1971, Fleiss' Kappa is a statistical measure designed to evaluate the degree of agreement among more than two annotators when assigning categorical ratings to a fixed number of items. It extends Cohen's Kappa, which is limited to two annotators, by accommodating multiple annotators and handling cases where different annotators evaluate different subsets of items. Compared to alternative reliability measures such as Krippendorff's Alpha or intra-class correlation coefficients (ICCs), Fleiss' Kappa is advantageous in cases where annotators do not necessarily rate all subjects and where categorical rather than continuous data are analyzed.

Fleiss' Kappa (κ) quantifies inter-annotator agreement by comparing the observed agreement among annotators to the agreement expected by chance. The measure accounts for variations in category distributions and is particularly useful in scenarios where ratings are nominal (categorical without a meaningful order). The value of Fleiss' Kappa ranges from -1 to 1, where higher values indicate stronger agreement: a κ of 1 signifies perfect agreement, 0 indicates agreement equal to chance, and negative values suggest disagreement greater than what would be expected randomly.

Let S be the set of samples, A be the set of annotators and k be the number of categories. The formula for Fleiss' Kappa is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.1}$$

where P_o is the observed agreement, calculated as:

$$P_o = \frac{1}{|S|} \sum_{s \in S} \left(\frac{\sum_{j=1}^{k} n_{sj}(n_{sj} - 1)}{n_s(n_s - 1)} \right) \tag{4.2}$$

and P_e is the expected agreement by chance, given by:

$$P_e = \sum_{j=1}^{k} p_j^2 \tag{4.3}$$

where n_{sj} is the number of annotators who assigned category j to sample s, n_s is the total number of ratings for sample s and p_j is the proportion of all assignments to category j across all samples:

$$p_j = \frac{1}{|S|} \sum_{s \in S} \frac{n_{sj}}{n_s} \tag{4.4}$$

Note that in this formulation κ is undefined if there are samples without annotations or just one annotation, since eq. 4.2 would produce a division by zero. Therefore, these samples are removed from the dataset before computing κ.

4.1.2 A novel inter-annotator agreement metric based on F1-score

As described in the introduction, despite being a robust metric for measuring inter-annotator agreement, Fleiss' Kappa is not comparable to classification metrics widely used in machine learning. Therefore, it does not enable us to determine whether a given metric score is satisfactory, that is, if it is comparable to the performance of annotators. To fill this gap, we define a novel metric based on the F1-score. We remind that the F1-score is defined as the harmonic average between precision and recall, where precision is the fraction of correctly identified instances among all instances predicted

as positive, and recall is the fraction of correctly identified instances among all actual positive instances. The F1-score is a versatile metric that can be applied to various classification scenarios, including binary, multi-class, and multi-label classification. The F1-score can be used to evaluate the agreement between two annotators, by considering one annotator as the gold standard and the other one as the predictor. However, there is no trivial way to accommodate more than two annotators and to manage a different set of annotators per sample. Moreover, the F1-score tends to overestimate the performances when the categories are balanced or unbalanced towards the negative class, as we show in our experimental analysis.

We start by describing our metric for binary data. The main idea is to use the F1-score for classification to evaluate the performance of each annotator and then normalize the result in a manner similar to how Fleiss' Kappa operates. In this way, we take advantage of both sides: we can support any kind of settings including multi-label (more than one choice per annotator), focus on a specific class (e.g., positive samples), and, at the same time, have a clear understanding of how much our performance stands out from chance. We consider a set of annotators A and a set of samples S. Each annotator $a \in A$ covers a subset of samples that we denote with S_a. We also denote as $f(a, s)$ the (0 or 1) value given by annotator a to sample s and with A_s the set of annotators that have annotated sample s.

To compute F1 we consider one annotator at a time and compute their classification performance considering the other annotations as the gold standard[1]. We adapt the precision and recall concepts to handle disagreement among the "gold" annotators by replacing the true values with the average "gold" annotations $Avg_{\neg a}$. We discard from S the set of samples s

[1] an equivalent alternative is compare each pairs of annotators, then averaging

such that $|A_s| \leq 1$, since no agreement can be computed on them. We then update the sets S_a accordingly. Given a sample s we define:

$$Avg_{\neg a}(s) = \frac{1}{|A_s| - 1} \sum_{a' \in A_s/a} f(a', s) \tag{4.5}$$

F1 is computed for each annotator $a \in A$. We define it in terms of (adapted) precision (p_a) and recall (r_a) .

$$p_a = \frac{\sum_{s \in S_a} \min(Avg_{\neg a}(s), f(a, s))}{\sum_{s \in S_a} f(s, a)} \tag{4.6}$$

$$r_a = \frac{\sum_{s \in S_a} \min(Avg_{\neg a}(s), f(a, s))}{\sum_{s \in S_a} Avg_{\neg a}(s)} \tag{4.7}$$

$$\mathrm{F1}_a = 2 \cdot \frac{p_a \cdot r_a}{p_a + r_a} \tag{4.8}$$

We then average among all annotators:

$$\mathrm{F1} = \frac{1}{\sum_{a \in A} |S_a|} \sum_{a \in A} |S_a| \cdot \mathrm{F1}_a \tag{4.9}$$

For multi-class and multi-labels we consider each class separately and then compute the weighted average among all classes. Therefore, we substitute f with f_c where $f_c(a, s)$ is 1 if the annotator a assigns class $c \in \mathbf{C}$ (set of classes) to sample s, zero otherwise. In this case, F1 is computed as:

$$multiclass\_\mathrm{F1} = \frac{\sum_{c \in \mathbf{C}} w_c \cdot \mathrm{F1}^c}{\sum_{c \in \mathbf{C}} w_c} \tag{4.10}$$

where w_c is computed as:

$$w_c = \sum_{a \in A} \sum_{s \in S_a} Avg^c_{\neg a}(s) \tag{4.11}$$

with $\mathrm{F1}^c$ and $Avg^c_{\neg a}$ computed as F1 and $Avg_{\neg a}$, but replacing f with f_c. As for Fleiss' Kappa we normalize F1 by computing its expected value, i.e., the value we would obtain by chance. To do so we define the expected precision $\hat{p}_a$ and expected recall $\hat{r}_a$ as:

$$\hat{p}_a = \frac{\sum_{s \in S_a} Avg_{\neg a}(s)}{|S_a|} \tag{4.12}$$

$$\hat{r}_a = \frac{\sum_{s \in S_a} f(a, s)}{|S_a|} \tag{4.13}$$

and compute the remaining expected metrics $\hat{\mathrm{F1}}_a$, $\hat{\mathrm{F1}}$ etc. accordingly. Finally we define our metric F1-kappa as:

$$F1 - \kappa = \frac{F1 - \hat{F}1}{1 - \hat{F}1} \tag{4.14}$$

As for Fleiss' κ, the value of F1-κ ranges from -1 to 1, where higher values indicate stronger agreement: a F1-κ of 1 signifies perfect agreement, 0 indicates agreement equal to chance, and negative values suggest disagreement greater than what would be expected randomly.

Now we describe how to evaluate the performance of a classifier in a way that is comparable with F1-κ. The proposed F1-κ metric can be interpreted as a measure of the average human performance in annotating a dataset and, therefore, serves as a reference for classification performance. The classifier evaluation is done by computing the same F1-κ with the only difference of substituting the annotations $f(a, s)$ of annotator a in eq. 4.6 and eq. 4.7 with the results of the classifier. We still need to exclude one annotator at a time from the gold standard to ensure that the evaluation is performed under the same conditions, thereby guaranteeing that the result can be directly compared with the inter-annotator agreement F1-κ.

4.2 Results and Evaluation

We conduct an experimental analysis to assess the effectiveness of the F1-κ metric in both binary and multi-label scenarios. In Section 4.2.1, we compare the performance of our metric with the adapted Fleiss' Kappa (cf. Sect. 4.1.1), aiming to assess their consistency. We conduct our evaluation on different synthetic datasets designed to simulate various binary classification scenarios. In Section 4.2.2, we show how F1-κ can be applied on a real-world multi-label classification scenario, specifically addressing the highly subjective domain of moral value identification.

4.2.1 Binary classification task

To assess the robustness of our metric, we construct synthetic datasets that emulate a binary classification task under varying annotation conditions. These datasets are designed to reflect real-world scenarios characterized by various degrees of agreement, label uncertainty, and varying rates of missing annotations. The data generation process begins with the creation of a reference truth vector, where binary labels are assigned to samples based on a predefined probability distribution (probability of positive). Annotators then provide labels with a certain probability of agreement with the reference truth. In cases of not agreement, the annotator generates a random label with the same probability distribution of the reference. Additionally, a proportion of annotations is intentionally omitted to simulate missing data. We compare F1-κ with the modified version of Fleiss' kappa (cf. Sect. 4.1.1), which enables handling missing values.

Figures 4.1 and 4.2 present the results of our proposed metric, F1-κ, alongside Fleiss' Kappa on a synthetic dataset of 500 samples annotated by

five different annotators. Additionally, Figure 6.2 reports the performance of F1-score (Eq. 4.9), which lacks the normalization process introduced in F1-κ (Eq. 4.14). This comparison highlights the impact of our balancing mechanism, demonstrating the independence of the F1-κ metric from the percentage of positive cases under varying conditions. Unlike the conventional F1-score, which is influenced by the distribution of positive and negative instances, the F1-κ provides a more robust assessment of inter-annotator agreement, unaffected by class imbalance. We evaluate metrics on datasets with varying agreement probabilities (ranging from 0.1 to 1.0) and probabilities of positive labels set at 0.1, 0.5, and 0.9. Additionally, we report the confidence intervals calculated over 10 independent runs to account for variability in the results.

Figures 4.2 and Figure 4.1 show the F1-κ and Fleiss' Kappa scores across varying agreement probabilities and with different probability of positive values ($p = 0.3$, $p = 0.5$ and $p = 0.9$). Both metrics demonstrate a similar, generally linear, increase in agreement as the threshold is raised. Specifically, at a positive class probability of 0.1, both F1-κ and Fleiss' Kappa initiate near zero and incrementally approach values of approximately 0.3 and 0.8 at thresholds of 0.5 and 0.9, respectively, ultimately converging towards perfect agreement (1.0). Moreover, values are independent of the probability of positive values, showing stability (i.e. independence from the percentage of positive cases) at different degrees sparsity in the data. Notably, the importance of the balancing mechanism inherent in F1-κ becomes evident when contrasted with the F1-score without normalization by expected value (Figure 6.2). The F1-score exhibits a marked sensitivity to the probability distribution of classes. At low agreement probabilities (i.e., 0.1), the weighted F1-score closely mirrors the distribution of positive instances

within the synthetic dataset. That is, F1-score values approximate the positive probability (e.g., ∼0.1 for p = 0.1, ∼0.5 for p = 0.5, and ∼0.9 for p = 0.9). This bias decreases with increasing agreement probabilities as the weighted F1-score asymptotically approaches 1.0. The behaviour depends on the fact that when the percentage of positive samples is high the chance of observing a true positive is high even in random data. This demonstrates how the balancing mechanism by expected value empowers F1-κ to provide a robust and reliable measure of inter-annotator agreement, exhibiting a behaviour comparable to Fleiss' Kappa in terms of resilience to class distribution imbalances within the dataset.

This trend is also observed in the presence of missing data. Figures 4.5, 4.4, and 4.6 illustrate the performance of F1-κ, Fleiss' Kappa, and the unbalanced F1-score, respectively, under varying percentages of missing data (0.1, 0.5, and 0.9) and agreement probabilities, with the probability of positive labels fixed at $p = 0.5$. The behaviour of Fleiss' Kappa and F1-κ remains highly similar, replicating the trends observed in the absence of missing data (Figures 4.2 and 4.4). However, the standard deviation of both metrics increases notably with a missing data rate of 0.9. Furthermore, at a low agreement probability of 0.1, both Fleiss' Kappa and F1-κ exhibit slightly lower values compared to the complete-data scenario (below the 0 threshold). This is due to the reduction in labeled data at high missing data rates. Comparison with the unbalanced F1-score (Figure 4.6) reveals results consistent with those presented in Figure 3. Specifically, the unbalanced F1-score initiates near 0.5, reflecting the balanced class distribution in the synthetic dataset, and converges towards 1.0 with increasing agreement probabilities.

Figure 4.1: Fleiss' κ with complete annotations

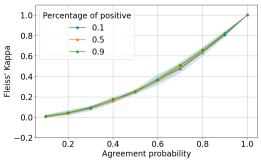


Figure 4.2: F1-κ with complete annotations

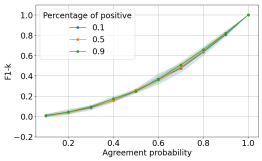


Figure 4.3: F1-score with complete annotations

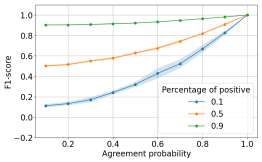


Figure 4.4: Fleiss' κ with missing annotations ($p = 0.5$)

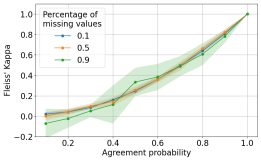


Figure 4.5: F1-κ with missing annotations ($p = 0.5$)

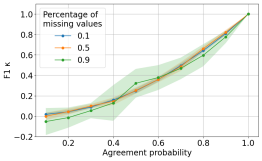
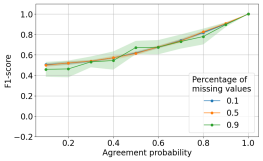


Figure 4.6: F1 with missing annotations ($p = 0.5$)



4.2.2 Case study on moral values identification

To assess the performance of our metric on real-world multi-label data, we present a case study on moral value classification, a highly subjective task. Specifically, we examine the work of Bulla et al. [53] (cf. Chapter 3), which assesses the performance of several LLMs in detecting moral values within the MFRC [337]. In Bulla et al. GPT-4 achieves the highest overall performance, surpassing human annotators in all moral dimensions. However, the model exhibits challenges in distinguishing between dyads that are inherently more ambiguous in textual contexts (e.g., Authority/Subversion). To ensure a fair comparison between human and AI performance, the study implements an evaluation methodology that mitigates biases introduced by annotation aggregation, which is often susceptible to low inter-annotator agreement. Specifically, the authors adopt an F1-score-based approach, wherein model predictions are compared against aggregate annotations from the MFRC while systematically excluding one annotator at a time. This methodology enables a direct and equitable comparison between human an-

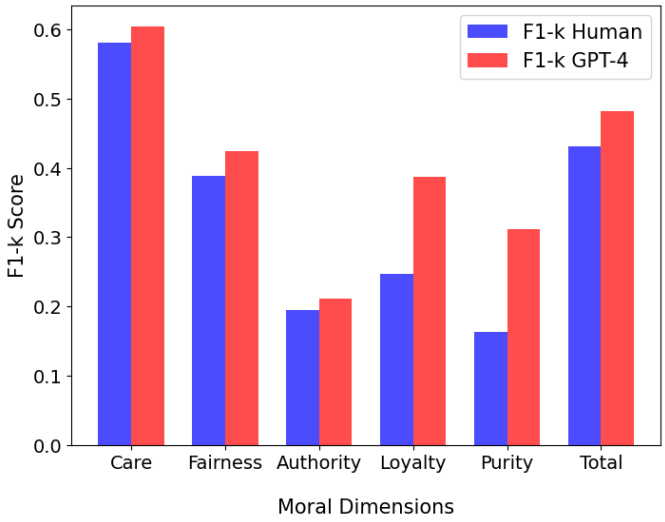
notators and model performance, with final results obtained by averaging across all annotators. The gold-standard annotations are determined by applying a threshold of 0.50 to individual annotation scores.

In our analysis, we compute the F1-κ metric for each moral dyad, enabling a direct comparison and assessing the consistency between GPT-4 predictions and annotator-assigned labels. We do not compare the F1-κ metric with Fleiss' Kappa, as the latter is not directly comparable to the former. Fleiss' Kappa measures inter-annotator agreement, whereas the F1-κ metric incorporates both precision and recall, which introduces a distinct evaluative dimension that cannot be directly aligned with Fleiss' Kappa in this context. This approach offers insights into our metric's adequacy in a real-world multi-label scenario, providing a setting for evaluating its effectiveness. To achieve this, we apply the F1-κ metric to estimate the inter-annotator agreement for each individual dyad as described in Section 4.1.2. We then compute the analogous metric by substituting the annotations of the reference annotator with the model's predictions, as described in the end of Section 4.1.2.

Figure 4.7 presents the F1-κ scores for all moral dyads on the MFRC dataset, comparing the predictions of GPT-4 with those of human annotators. For this comparison, we select the highest-performer model from Bulla et al., specifically the GPT-4 outcomes in a double-prompt scenario. In this scenario, GPT-4 is first prompted to determine whether the input conveys moral content and, if so, is prompted again to identify which moral dyads are present in the text through a two-step inference process. As depicted in figure (Fig. 4.7), our results are consistent with those reported in Bulla et al., where GPT-4 outperforms human annotators in all moral dyads, achieving the highest performance in Care and Fairness dyads. On average across all

moral dyads, GPT-4 outperforms human annotators with an improvement of 5 percentage points in terms of F1-κ score (i.e. 0.48 vs 0.43). Specifically, GPT-4 achieves an F1-κ of 0.61 and 0.42 for Care and Fairness, compared to 0.58 and 0.39 for human annotations. However, for the more opaque dyads (i.e. Authority, Loyalty, and Purity), GPT-4's F1-κ scores are 0.21, 0.39, and 0.31, respectively, in contrast to 0.19, 0.25, and 0.16 for human annotators. The F1-κ values in Figure 4.7 are slightly lower than the F1 scores reported in Bulla et al., primarily due to the normalization process applied in our F1-κ metric (cf. Sect.4.1.2), which adjusts the results according to the expected F1-score, calculated based on the label distribution across the entire annotation corpus. This normalization enables adapting the scale such that a random level of agreement is lowered to 0. In contrast, the standard F1 score does not account for agreement by chance. Notably, the comparison with human annotations is crucial for accurately evaluating model performance in this inherently subjective task. By performing a direct comparison, we establish a scale that contextualizes the performance of both models and human annotators in relation to the subjectivity of the task. This comparative framework allows for a more nuanced interpretation of the results, positioning model and human performance with respect to the chance threshold and providing deeper insights into the model's alignment with human moral judgment.

Figure 4.7: The proposed F1-κ metric enables directly comparing inter-annotator agreement with the model's performance. In moral value detection, the model outperforms human annotators. This result is consistent with earlier findings.



5 Grammar-Constrained Natural Language Generation

Within the Human-Centered AI paradigm, aligning AI outputs with predefined, human-understandable structures is crucial for transparency, trust, and usability (cf. Sect. 1.2.3 and 2.2.3). This motivates the need to constrain state-of-the-art generative models so that they not only produce high-quality text but also deliver controllable and explainable outputs. To this end, we present GRAMMAR-LLM [339], a system that integrates formal grammar constraints into LLMs to enforce both syntactic and semantic validity during generation.

As discussed in Section 2.2.3, the demand for controllable models has become increasingly pressing, as such models are essential for fostering human–AI trust and ensuring the reliability of intelligent systems. Despite their impressive performance in tasks such as machine translation, text summarization, and open-domain generation [241], LLMs inherently lack mechanisms to constrain their outputs according to predefined taxonomies [121]. Many real-world applications require fine-grained control over generated text, such as enforcing domain-specific rules or syntactic regularities. Without such control, LLMs often fail in tasks that demand strict lexical or structural conformity, including classification or question answering with predefined options.

Moreover, since LLMs are highly dependent on their training data, their performance degrades when generating text in languages or modalities that lie outside the training distribution. This issue is especially severe in low-resource scenarios, such as underrepresented or ancient languages, where insufficient data prevents the models from capturing necessary structures. Ensuring that LMs adhere to predefined structural and lexical constraints, therefore, remains a significant challenge.

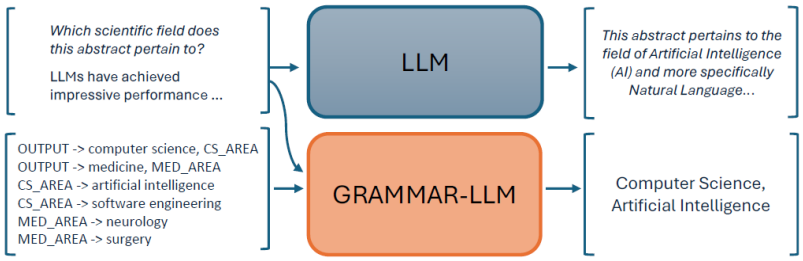


Figure 5.1: GRAMMAR-LLM enforces LLMs to adhere to a predefined grammar, thus avoiding verbose answers when a structured output is required.

To address these issues, GRAMMAR-LLM incorporates formal grammars directly into the decoding process (Fig.5.1). Unlike post-hoc validation or heuristic filtering, our method treats text generation as an automaton-driven process, ensuring compliance with a predefined Context-Free Grammar (CFG) [149].

However, integrating grammatical constraints poses challenges due to the complexity of formal grammars and the trade-off between expressiveness and computational efficiency. Highly expressive grammars provide strong control but may become computationally intractable in autoregressive settings. Restricting to linear-time grammars, such as LL(1) [6], alleviates this, but designing LL(1) grammars for tokenized text is difficult, especially with sub-

word tokenization. To overcome this, we introduce LL(prefix), a novel notation for LL(1) grammars that enables grammar definition independently of tokenization. A PushDown Automaton (PDA) [88], automatically derived from the input grammar, constrains the model to generate only valid tokens.

Our framework is compatible with different LM architectures, including GPT- and BART-based models, and guarantees outputs that strictly conform to predefined grammatical structures.

We evaluate GRAMMAR-LLM across three domains: hierarchical classification [324], sign language translation [246], and semantic parsing [172]. Each requires structured outputs aligned with taxonomies or specialized lexicons that can be effectively represented through CFGs. Hierarchical classification demands outputs that preserve taxonomy paths; sign language translation requires bridging modality gaps in low-resource conditions; and semantic parsing necessitates adherence to strict semantic schemas. To examine the influence of model scale and architecture, we test GRAMMAR-LLM on LLaMA-3 variants (1B, 8B, 70B) [96] and a BART-based model [32].

Across all tasks, GRAMMAR-LLM consistently outperforms baselines, achieving substantial gains in classification accuracy, translation quality, and compliance with structural constraints. These results underscore the value of integrating formal grammars into LLMs, especially in applications requiring structured and highly precise outputs.

In Section 5.1, we detail the GRAMMAR-LLM framework. Section 5.2 presents experimental results, and Section 5.3 concludes with a discussion of limitations. Chapter 2.2.3 presents an in-depth discussion about the state of the art of this field.

5.1 A Novel Framework for Grammar- Constrained LLM Decoding

Our method builds upon the literature on context-free grammars to design an efficient mechanism that enforces LLMs to generate grammar-compliant output. We first introduce the literature on LL(1) grammars and their implementation by means of pushdown automata, which enable left-to-right processing in linear time and are therefore suitable for integration with LLMs. Then, we introduce a novel and more expressive formalization, named LL(prefix), which is well-suited for representing sequences of tokens in a user-friendly way, and give an algorithm to transform LL(prefix) grammars into LL(1) grammars. Eventually, we show how to integrate a deterministic pushdown automaton with a transformer decoder to enforce text generation that adheres to the input LL(prefix) grammar. An overview of our pipeline is shown in Figure 5.2.

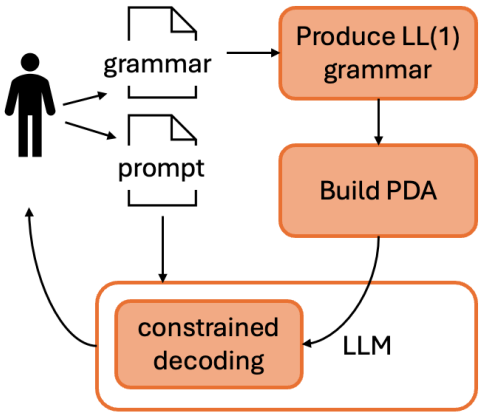


Figure 5.2: Overview of our GRAMMAR-LLM pipeline.

5.1.1 Background

In this section we briefly introduce LL(1) grammars and its use for verification by means of pushdown automata.

We begin with the definitions of CFGs and LL(1) grammars [149].

Definition 1. *A grammar $G = (\Sigma, N, P, S)$ is context-free if all its production are of the form $A \to \alpha$ with $A \in N$ and $\alpha \in (\Sigma \cup N)^*$.*

A significant subclass of the family of context-free grammars is the set of LL(1) grammars. The definition of LL(1) grammar is based on two functions $FIRST$ and $FOLLOW$ associated with the grammar. For any $\alpha \in (\Sigma \cup N)^*$, $FIRST(\alpha)$ is the set of terminals that begin the strings derived from α. If $\alpha \to \epsilon$ then $\epsilon \in FIRST(\alpha)$.

For any $A \in N$, $FOLLOW(A)$ is the set of terminals a that can appear immediately to the right of A in some sentential form (i.e. the set of terminals a such that there exists a derivation of the form $S \Rightarrow^* \alpha A a \beta$). If A can be the rightmost symbol in some sentential form, the \$ is in $FOLLOW(A)$, where \$ is an endmarker symbol.

Definition 2. *A context-free grammar $G = (\Sigma, N, P, S)$ is LL(1) if whenever $A \to \alpha \mid \beta$ are two distinct productions of G, the following conditions hold:*

1) For no terminal $a \in \Sigma$, both α and β derive a string beginning with a (i.e. no terminal $a \in \Sigma$ is in $FIRST(\alpha) \cap FIRST(\beta)$)

2) At most one of α and β can derive ϵ

3) If $\beta \Rightarrow^* \epsilon$, then α does not derive any string beginning with a terminal in $FOLLOW(A)$ (i.e. if $\alpha \Rightarrow^* a\alpha_1$ then $a \notin FOLLOW(A)$).

Languages accepted by LL(1) grammars can be recognized using pushdown automata (PDAs). We now give a formal definition of a PDA, followed

by a description of the algorithm to construct the automaton.

A *pushdown automaton*, PDA, for short, is a system $M = (Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$ where Q is a finite set of *states*, Σ is the *input alphabet*, Γ is the *stack alphabet*, $q_0 \in Q$ is the *initial state*, $Z_0 \in \Gamma$ is the *start symbol*, $F \subseteq Q$ is the set of *final sates* and δ, the *transition function*, is a mapping from $Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma$ to finite subsets of $Q \times \Gamma^*$.

The interpretation of $\delta(q, a, Z) = \{(p_1, \gamma_1), \ldots (p_m, \gamma_m)\}$, with $q, p_1 \ldots p_m \in Q$, $a \in \Sigma$, $Z \in \Gamma$ and $\gamma_1 \ldots \gamma_m \in \Gamma^*$, is that the PDA in the state q, with input symbol a and symbol Z at the top of the stack, can for any i enter state p_i, replace symbol Z by γ_i and advance the input head one symbol. The interpretation of $\delta(q, \epsilon, Z) = \{(p_1, \gamma_1), \ldots (p_m, \gamma_m)\}$ is that the PDA in the state q, independent of the input symbol being scanned and symbol Z at the top of the stack, can enter state p_i, replace symbol Z by γ_i. In this case the input head is not advanced. For a PDA the language accepted by empty stack is defined in terms of *instantaneous descriptions*, ID for short, that formally describe the configurations of a PDA at a given instant: an ID is a triple (q, w, γ) where $q \in Q$ is the current state, $w \in \Sigma$ is the "unexpended input" and $\gamma \in \Gamma^*$ is the stack content. Given two ID, I_1 and I_2, we write $I_1 \vdash^* I_2$ if I_2 can be reached from I_1 after 0 or more steps. Then, for a PDA M, the language accepted by empty stack is $N(M) = \{w \in \Sigma^* \mid (q_0, w, Z_0) \vdash^* (p, \epsilon, \epsilon) \text{ for some } p \in Q\}$. A PDA is said *deterministic* if at most one move is possible from any ID.

It is known, that any language generated by a context-free grammar can be accepted, by empty stack, by a PDA M [149]; but, here, since we consider LL(1) context-free grammars and since we need deterministic PDA, we present a different algorithm for the construction of M (see Algorithm 1) inspired by the construction of a Parsing Table for an LL(1) grammar [6].

Algorithm 1 Construction of a PDA, M, from an LL(1) grammar G

Require: An LL(1) grammar $G = (\Sigma, N, P, S)$
Ensure: A pushdown automaton $M = (Q, \Sigma \cup \{\$\}, \Gamma, \delta, q_0, Z_0, \emptyset)$ equivalent to G
1: init. $Q \leftarrow \{q_a \mid a \in \Sigma\} \cup \{q_0, q_\$\}$
2: init. $\Gamma \leftarrow \Sigma \cup N \cup \{\$\}$
3: init. $Z_0 \leftarrow \$S$
4: For any $a \in \Sigma \cup \{\$\}$ and $Z \in N \cup \{\$\}$ do $\delta(q_0, a, Z) \leftarrow (q_a, Z)$
5: For any $a \in \Sigma \cup \{\$\}$ do $\delta(q_a, \epsilon, a) \leftarrow (q_0, \epsilon)$
6: **for all** productions $A \to \alpha$ in P **do**
7: For each terminal $a \in FIRST(\alpha)$ $\delta(q_a, \epsilon, A) \leftarrow (q_a, \alpha)$
8: If $\epsilon \in FIRST(\alpha)$, for any $b \in FOLLOW(A)$, $\delta(q_b, \epsilon, A) \leftarrow (q_b, \alpha)$
9: If $\epsilon \in FIRST(\alpha)$ and $\$ \in FOLLOW(A)$, $\delta(q_\$, \epsilon, A) \leftarrow (q_\$, \alpha)$
10: **end for**
11: **Return** M

Theorem 5.1.1. *Algorithm 1, with an LL(1) grammar G in input, produces in linear time a deterministic PDA, M, that is equivalent to G, i.e. $N(M) = L(G)\$$, where $L(G)\$$ is the set of strings in $L(G)$ followed by an endmarker* \$.

Note that, since G is LL(1) then $|\delta(q, a, A)| \leq 1$, for any $a \in \Sigma \cup \{\$, \epsilon\}$ and $A \in \Gamma$. Moreover, if $\delta(q, \epsilon, A) \neq \emptyset$, then $\delta(q, a, A) = \emptyset$, for any $a \in \Sigma \cup \{\$\}$. Hence M is a deterministic PDA. At last, one can show, that $N(M) = L(G)\$$.

5.1.2 A Novel Expressive Grammar Notation for Constraining LLM Output

Despite implementable with pushdown automata, LL(1) grammars exhibit limited expressiveness when applied to LLMs. This limitation arises because grammars are traditionally defined in terms of symbols, which can correspond to words or characters. However, the fundamental semantic unit for LLMs is the token, which often represents subword segments rather than entire words. Using tokens as terminal symbols in a grammar would make defining rules cumbersome for users, making this approach impracti-

cal. While words or entire sentences in a grammar can be easily converted into sequences of tokens, this conversion may violate the constraints of an LL(1) grammar. To illustrate this issue, consider the following grammar:

$S \rightarrow$ uncertain S | undefined S | ϵ

If we consider the words "uncertain" and "undefined" as terminals, this grammar is LL(1). However if we adopt subword tokenization where the prefix "un" is separated by the rest of the word, the grammar turns into:

$S \rightarrow$ un certain S | un defined S | ϵ

which is violates condition 1) of LL(1) grammars (Def. 2) since the RHS of two different rules with the same LHS begin with the same terminal "un".

To overcome this limitation, we define a class of grammars, – which we show to be equivalent in generative capacity to LL(1) – named $LL(prefix)$, which allows multiple rules to share sequences of symbols in the initial portion of the RHS of production rules. The definition generalizes traditional $LL(1)$ grammars by accommodating prefix-sharing structures. We then demonstrate that any $LL(prefix)$ grammar can be transformed into an equivalent $LL(1)$ grammar in linear time and space (Theorem 5.1.2 below). This transformation ensures that $LL(prefix)$ grammars remain efficiently parseable and can be verified using pushdown automata, maintaining integrability with LLMs, as we show in Sect. 5.1.3.

Definition 3. *A context-free grammar $G = (\Sigma, N, P, S)$ is LL(prefix) if whenever $A \rightarrow \omega\ \alpha \mid \omega\ \beta$ are two distinct productions of G with $\omega \in \Sigma^*$ and α and β starting with a different symbol, the following conditions hold:*

1) For no terminal $a \in \Sigma$, both α and β derive a string beginning with a (i.e. no terminal $a \in \Sigma$ is in $FIRST(\alpha) \cap FIRST(\beta)$)

2) At most one of α and β can derive ϵ

3) If $\beta \Rightarrow^* \epsilon$, then α does not derive any string beginning with a terminal

in $FOLLOW(A)$ *(i.e. if* $\alpha \Rightarrow^* a\alpha_1$ *then* $a \notin FOLLOW(A)$*).*

In short, with respect to LL(1) grammars, Definition 3 admits prefixes of terminals of any length to be shared among different rules with the same LHS.

We give an algorithm to transform an LL(prefix) grammar into an LL(1) grammar and prove that it correctly generates an equivalent LL(1) grammar in linear time and space.

Algorithm 2 Transforming an LL(prefix) grammar into an LL(1) grammar

Require: An LL(prefix) grammar $G = (\Sigma, N, P, S)$
Ensure: An equivalent LL(1) grammar G'
1: **while** there are pairs of productions of the type $A \to a\ \alpha$ and $A \to a\ \beta$ for some $A \in N$ and $a \in \Sigma$ **do**
2: Take all productions of the kind $A \to a\ \alpha_i$ for every $\alpha_i \in (\Sigma \cup N)^*$ and substitute them with: $A \to a\ B$ and $B \to \alpha_i$ for every α_i, where B is a newly generated non-terminal
3: Add B to N
4: **end while**
5: **Return** $G' = (\Sigma, N, P, S)$

Theorem 5.1.2. *Algorithm 2 produces an LL(1) grammar* G' *that is equivalent to the input LL(prefix) grammar* G *with time and space complexity* $O(n \cdot m)$ *where* n *is the number of production rules and* m *is the maximum length of production rules.*

The proof of the theorem is available in the supplemental material.

5.1.3 Grammatically-Compliant Text Generation

We show how to use a pushdown automaton to enforce a Transformer decoder to generate valid text conforming to an LL(prefix) grammar. As described in Section 5.1.2, any LL(prefix) grammar can be transformed into an equivalent LL(1) grammar. In turn, an LL(1) grammar can be recognized by a pushdown automaton, as discussed in Section 5.1.1. In the following,

we describe how to integrate a pushdown automaton into the generation process of an LLM to ensure grammar compliance.

LLMs generate tokens one by one following an autoregressive decoding approach, where each new token is predicted based on input and previously generated tokens. To do so, the output from the last decoder layer is transformed to the vocabulary space using a linear projection and a softmax function:

$$\begin{aligned} P(y_t \mid x_1, \ldots, x_n, y_1, \ldots, y_{t-1}) = \\ = \text{softmax}(W_o H_t + b_o) \end{aligned} \tag{5.1}$$

where H_t is the output from the last decoder layer and W_o, b_o are learnable parameters. The process repeats iteratively until a stopping criterion is met.

We show that a pushdown automaton can be integrated with the decoding process to force the LLM to follow the input grammar. Our automaton considers tokens as terminal symbols. The generation process is described in Algorithm 3.

Algorithm 3 Autoregressive Token Generation

Require: Initial sequence of tokens $X = \{x_1, x_2, \ldots, x_n\}$, pushdown automaton $M = (Q, \Sigma, \Gamma, \delta, q_0, \gamma_0, F)$, model parameters θ
Ensure: Generated token sequence $Y = \{y_1, y_2, \ldots, y_m\}$

```
init. Y ← X
init. PDA configuration (q, γ) ← (q_0, γ_0)
while γ is not empty do
    H_t ← Transformer(Y, θ)
    P(y_t | Y) ← softmax(W_o H_t + b_o) ∘ next_terminals((q, γ))
    sample next token y_t and append it to Y
    do transition (q, γ) ← δ(q, y_t, top(γ))
    while q ≠ q_0 do
        do transition (q, γ) ← δ(q, ϵ, top(γ))
    end while
end while
return Y
```

The differences with respect to standard decoding are in lines 2,3,5 and 7-10. Line 2 initializes the automaton. The exit condition in Line 3 is actually equivalent to standard decoding since the stack is empty when the end of sentence symbol \$ is generated. Line 5 generates the probability distribution for the next token considering forbidden tokens. next_terminals(q, γ) generates the set of valid tokens given the current automaton configuration and return them as a one-hot vector. The operator $\circ$ performs the element-wise product between two vectors. For LL(1) grammars, next_terminals$((q, \gamma))$ simply return $FIRST(\gamma)$. Lines 7-10 perform automaton transitions based on its current state, the newly generated token and the top of the stack. The first transition consumes the generated token and moves the automaton to a token-specific state. Transitions are performed until the automaton returns to the initial state, i.e. it is ready to accept another token.

5.2 Results and Evaluation

By enforcing strict compliance with formal grammars during text generation, GRAMMAR-LLM contributes to the development of AI systems that are not only accurate but also transparent and aligned with human expectations, supporting the overarching goals of Human-Centered AI To assess the effectiveness of our method, we conduct experiments across three distinct NLP tasks (i.e. hierarchical classification, sign language translation and semantic parsing), each chosen to test different aspects of constrained text generation. Specifically, in the hierarchical classification task [324], we assess the model's capability to generate text that conforms to predefined taxonomies, a critical requirement in structured data generation and classification scenarios. Sign language translation [246] refers to the translation from natural language to glosses, where sign language is encoded using gloss

notation. This notation aligns signs with words or phrases from a spoken language, serving as a structured, text-based annotation rather than a direct translation. This task poses a unique challenge due to the inherent modality gap between natural language and gloss-based representations, as well as the scarcity of high-quality parallel data. We employ this case study to highlight the effectiveness of GRAMMAR-LLM in low-resource and modality-specific contexts, where conventional LLMs often struggle with distributional mismatches. The semantic parsing [172] demands specific alignment between generated text and predefined semantic schemas, making it a robust benchmark for assessing GRAMMAR-LLM's ability to enforce formal grammatical constraints. This task underscores the model's adaptability to fine-tuned structured generation tasks, ensuring that outputs maintain both semantic accuracy and syntactic coherence[1].

Table 5.1 presents the results for the hierarchical classification task in terms of micro F1 score. We test our method on the Web of Science (WoS) dataset [192], which consists of approximately 50,000 research abstracts annotated with a two-level taxonomy consisting of 7 and 134 labels at Levels 1 and 2, respectively. For evaluation, we select 2,000 instances using a stratified sampling strategy at Level 1 to ensure proportional representation, while employing random sampling at Level 2. In all cases, the prompt explicitly specifies the full taxonomy and the expected output format. We assess the performance of the LLaMA-3[2] model in three configurations (1B, 8B, and 70B) across zero-shot, one-shot, and few-shot scenarios. In the few-shot set-

[1]Further details on the prompts used in the experiments (Tables 5.1 and Table 5.3) and the grammars implemented for the three case studies are provided in the supplementary materials.

[2]We chose LLama-3 as a representative LLM since it is one of the most popular open-source model and it achieves competitive performances with respect to commercial models. Most commercial models (e.g. GPT-4) cannot be employed since we cannot intervene in their decoding process.

ting, we provide 10 examples, selected using the same sampling criteria as the test set. As a baseline, we compare these models with and without our GRAMMAR-LLM (Sect. 5.1) approach (which we name G-LlaMa) to quantify its impact on classification performance. We employ a simple grammar that forces the model to first generate a Level 1 class, then generate a Level 2 class that is compatible with the Level 1 class.

As shown in Table 5.1, G-LLaMA outperforms unconstrained LLaMA models across all experimental settings. The G-LLaMa model achieves the best results in the zero-shot scenario, with F1 scores of 0.734 and 0.504 at the first and second taxonomy levels, respectively. An analysis of Level 1 classification reveals that the most significant improvements occur in the smallest model (i.e. LLaMA-1B). In this case, the constrained model achieves an F1 score improvement of 30 and 28 percentage points over the unconstrained baseline in the zero-shot and one-shot configurations, respectively, showing comparable performance in the few-shot settings. These findings suggest that larger models with more parameters benefit less from the constraints at the first taxonomy level, as it is easier to classify due to broader, distinct categories. In contrast, smaller models benefit more from the grammar-based module, which enables them to generate valid, concise outputs and improves accuracy, as demonstrated by LLaMA-1B's zero-shot and one-shot performance. At a more granular taxonomy level (i.e. Level 2), the constrained models exhibit more pronounced improvements over the baseline models across all configuration settings. Specifically, G-LlaMa-1B outperforms the unconstrained model by 12, 16, and 8 percentage points in the zero-shot, one-shot, and few-shot settings, respectively. A similar trend is observed for LLaMA-8B and LLaMA-70B, where the grammar-based models yield an F1 score improvement of 5, 4, and 6 percentage points for the Llama-8B,

and 4, 2, and 2 points for Llama-70B, respectively, across the zero-, one- and few-shot settings. At the second-level taxonomy, the grammar-based constraints lead to more significant improvements, mitigating the generative models' verbosity. This ensures adherence to a valid hierarchical path and avoids the generation of invalid taxonomy labels[3].

Notably, the percentage of outputs conform to the requirements (i.e. referred to as validity in Table) closely follows the improvements in F1 score. The constrained models consistently achieve a 100% validity rate across all configurations, which indicate outputs that fully comply with the target taxonomy. In contrast, unconstrained models yield substantially lower validity rates, with performance varying significantly across model sizes. Specifically, LLaMA-1B, LLaMA-8B, and LLaMA-70B achieve best validity rates of 27%, 69%, and 85%, respectively, showing an improvement as the number of examples in the prompt increases. Nevertheless, none of the unconstrained models reach full validity, adversely affecting their overall performance.

To assess the types of errors that our approach mitigates, Table 5.2 presents the main error categories produced by unconstrained models in the classification task across all prompt configurations. We classify three mutually exclusive main categories of invalid outputs: invalid taxonomy labels (i.e. cases where a label is appropriate at a general level but incorrectly assigned at a more specific level); incorrect number of labels (i.e. outputs with excessive hierarchical levels w.r.t. the predefined taxonomy, while remaining accurate at the first two levels), and others (i.e. verbose or free-form responses that do not adhere to the expected classification format). As shown in Table 5.2, the most frequent error across all models involve the

[3]As shown in Table 5.1, LLaMA-70B performs best in the zero-shot setting, likely due to its extensive parameterization. In contrast, few-shot prompting may limit generalization by overly anchoring the model to provided examples

Model	Zero-shot			One-shot			Few-shot		
	L1	**L2**	**Validity**	**L1**	**L2**	**Validity**	**L1**	**L2**	**Validity**
LlaMa-1B	.002	.001	0%	.313	.071	16%	.501	.092	27%
G-LlaMa-1B	**.308**	**.124**	100%	**.539**	**.231**	100%	**.504**	**.173**	100%
LlaMa-8B	.569	.229	62%	**.650**	.258	69%	**.680**	.255	62%
G-LlaMa-8B	**.575**	**.282**	100%	**.650**	**.298**	100%	**.680**	**.310**	100%
LlaMa-70B	.720	.464	80%	.740	.437	83%	.700	.362	85%
G-LlaMa-70B	**.734**	**.504**	100%	**.742**	**.456**	100%	**.702**	**.378**	100%

Table 5.1: Performance on hierarchical classification using the WoS dataset, evaluated in terms of micro F1 score for Level 1 (L1) and Level 2 (L2) classification, as well as the overall validity rate (%) for each model and configuration.

inclusion of invalid taxonomy labels or the omission of required ones. The only exception is observed in LLaMA-1B under the zero- and one-shot settings, where non-conforming outputs predominate. This depends on model's tendency towards verbosity in the absence of sufficient instructional context, often producing too long responses or generating content that is not related to the task. However, in the few-shot configuration, LLaMA-1B aligns with the error profile of larger models. The inclusion of additional examples in the prompt offers more explicit guidance, facilitating a deeper understanding of the task structure and substantially reducing the occurrence of unstructured outputs. LLaMA-8B and LLaMA-70B exhibit invalid taxonomy labels as the primary error, with LLaMA-70B producing virtually no other error category.

Table 5.3 provides an analysis of the performance of the constrained and unconstrained LlaMa models at varying model size (i.e. 1B, 8B and 70B) for the sign language translation task in a few-shot setting. We test models on 2,000 randomly sampled instances from the Synthetic English-ASL Gloss Parallel Corpus [251], which comprises 87,710 text-gloss pairs, generated through the application of syntactic transformation rules to English text. For each input sentence, we prompt models in few-shot by selecting

relevant examples based on cosine similarity [203]. Specifically, we retrieve the 30 most similar text-gloss pairs and the 50 most similar glosses from the training set. This approach ensures that the prompt remains contextually relevant to the input. For our method we adopt a simple grammar that enforces producing sequences of entries from the set of admissible glosses. We present results in terms of BLEU [257], ChrF [264], F1 score. In addition, we report the number of valid outputs in terms of percentage as in Table 5.1. BLEU measures the precision of n-grams between the predicted and reference translations, while ChrF focuses on character level to evaluate more granular alignment. The F1 score reflects the balance between precision and recall over the gloss vocabulary, treating the task as multi-label classification and hence discarding the order of glosses.

As shown in Table 5.3, model performance increases consistently with model size. This effect is more pronounced for smaller models (i.e. 1B and 8B), where the need for improvement is higher. We observe performance gains of 14, 5 and 4 percentage points in terms of F1 score for the 1B, 8B and 70B models, respectively, w.r.t. LlaMa baselines. These findings are consistent with the results reported for the classification task (cf. Table 5.1). We observe comparable trends in terms of ChrF2 and BLEU metrics. The analysis of output validity mirrors these trends, with all constrained models achieving 100% output validity, in contrast to significantly lower rates for their unconstrained counterparts. This discrepancy depends on the tendency of unconstained models to generate non-existent glosses, that is not possible with the structural constraints imposed in G-LLaMA models.

These findings highlight the impact of GRAMMAR-LLM in enhancing model performance, especially in translation tasks involving limited data and underrepresented domains. Specifically, the grammatical module con-

tributes to selecting glosses that consistently align with the vocabulary, thereby avoiding errors due to the generation of invalid labels.

Table 5.4 evaluates the performance of constrained and unconstrained BART-based models (AMR-BART-base - 139M - and AMR-BART-large - 406M [32]) for Abstract Meaning Representation (AMR) [35] parsing, a semantic framework representing sentence meaning as directed graphs. We test models on 100 randomly selected samples from the LDC dataset [47], which contains 39,260 sentence-AMR pairs. We compare models with and without the GRAMMAR-LLM module in terms of SMATCH F1 score [62], which measures structural similarity between AMR graphs, and report the percentage of valid graphs generated after fine-tuning and postprocessing, as AMR-BART uses a rule-based approach to rectify invalid outputs. As shown in Table 5.4, our G-AMRBART-base outperforms its unconstrained counterpart (AMRBART-base) by 2 percentage points. Meanwhile, G-AMRBART-large exhibits comparable performance to AMRBART-large, achieving a SMATCH F1 score of 0.789 compared to 0.79. The improved performance of the larger model can be attributed to its increased parameter count and extensive training data, enabling more effective fine-tuning and enhanced ability to process and learn the syntactic structure of AMR graphs. In this scenario, our method shows comparable performance with the unconstrained fine-tuned model, as no significant further enhancements are attainable. In contrast, the AMRBART-base model, constrained by its reduced parameter count, struggles to learn the syntactic rules, leading to invalid graph outputs in 6% of cases. Here, the integration of a grammar module ensures the generation of semantically valid outputs, thereby outperforming the baseline model.

5.3 Limitations

Our approach inherits the autoregressive nature of LLMs, where each token generated affects the probability distribution of subsequent outputs. This sequential dependency can lead to error propagation, as early-stage mistakes may significantly impact the quality of later tokens. Our approach could, in certain cases, amplify early-stage mistakes since validity constraints might intervene later on, preventing the generation of tokens which are necessary to complete a meaningful (though invalid) sentence. Techniques such as those proposed in Park et al. [260] mitigate early-stage errors by reshaping token distributions. We consider this work complementary to ours and advocate for the synergistic use of both approaches as future work.

Another limit is given by the expressiveness of our class of LL(prefix) grammars. Although more expressive than traditional LL(1) grammars, LL(prefix) grammars remain less expressive than CFGs. CFGs allow for grammar ambiguity and can model more complex syntactic structures, which may be necessary for certain advanced natural language generation tasks. This limitation means that while LL(prefix) grammars are well-suited for many structured generation tasks, they may struggle with highly recursive or nested language constructs that require the full power of CFGs. The downside of using the entire class of CFGs is that their languages cannot be accepted left-to-right by a deterministic PDA. Consequently, integrating them with LLMs [120] would introduce complex data structures and a super-linear overhead, with the exponent depending on the specific grammar [19].

Furthermore, our current implementation does not incorporate an end-to-end fine-tuning phase that integrates the grammar module into the training loop. As a result, our findings may not fully capture the potential behav-

ior of LLMs when fine-tuned on extensive datasets, particularly in scenarios where the grammar constraints are deeply embedded in the model's learning process. Exploring grammar-constrained fine-tuning remains an avenue for future research.

Model	Zero-shot			One-shot			Few-shot		
	Incorrect number of labels	**Invalid taxonomy labels**	**Others**	**Incorrect number of labels**	**Invalid taxonomy labels**	**Others**	**Incorrect number of labels**	**Invalid taxonomy labels**	**Others**
LlaMa 1B	0%	0%	100%	2%	30%	53%	0%	70%	3%
LlaMa 8B	2%	23%	13%	0%	28%	3%	0%	38%	0%
LlaMa 70B	0%	17%	2%	0%	17%	0%	0%	14%	0%

Table 5.2: Error analysis of unconstrained models across all prompt configurations in the classification task.

Model	BLEU	ChrF	F1	Validity
LlaMa-1B	0.31	0.72	0.65	9%
G-LlaMa-1B	**0.47**	**0.76**	**0.79**	100%
LlaMa-8B	0.70	0.90	0.89	39%
G-LlaMa-8B	**0.81**	**0.93**	**0.94**	100%
LlaMa-70B	0.79	0.93	0.92	49%
G-LlaMa-70B	**0.86**	**0.95**	**0.96**	100%

Table 5.3: Performance of the models on the text-to-gloss ASL translation task, evaluated on the Synthetic English-ASL Gloss Parallel Corpus in terms of BLEU, ChrF, and F1 scores. We report the overall validity rate (%) for each model and configuration.

Model	Smatch	Validity
AMRBART-base	0.559	94%
G-AMRBART-base	**0.577**	100%
AMRBART-large	0.789	100%
G-AMRBART-large	**0.790**	100%

Table 5.4: Performance of the models on the AMR semantic parsing task, evaluated on the LDC dataset in terms of Smatch metric.

6 Leveraging Large Language Models for Accurate Sign Language Translation in Low-Resource Scenarios

A central goal of this dissertation is to advance the HCAI paradigm by developing systems that are not only technically proficient but also aligned with human values, robust, controllable, and inclusive. While previous chapters have explored challenges related to moral value alignment (cf. Sect. 3), evaluation in subjective contexts (cf. Sect. 4), and controllable AI (cf. Sect. 5), an equally important dimension of HCAI is ensuring accessibility for underrepresented communities (cf. Sect. 1.3.1).

Among these communities, the Deaf community faces unique barriers in human–AI interaction, particularly in the domain of language technologies (cf. Sect. 2.3). Despite significant progress in speech and text-based NLP, sign languages remain underrepresented in both data resources and computational models.

This limitation stems from their reliance on training data dominated by widely spoken languages, which constrains their ability to understand, represent, and translate underrepresented linguistic systems. Among low-resource languages, sign language presents significant challenges. Sign languages, such as Italian Sign Language (LIS) and American Sign Language

(ASL), rely on a visual-spatial grammar system rather than spoken or written syntax. This linguistic structure, coupled with the scarcity of available training data, makes accurate translation more difficult to achieve. Furthermore, translating spoken language into sign language is under-explored in the current research. Addressing this challenge demands effective methods in data-scarce scenarios, leveraging external linguistic resources like specialized vocabularies and lexicons to produce coherent and explainable translations.

Sign Language Translation (SLT) is usually approached by separating the graphical part (computer vision or video generation) from the language part (translation) and use an intermediate language for encoding the sign language. One common representation is based on glosses. Despite its simplicity, gloss-based representations tend to oversimplify the linguistic complexity of sign languages. To overcome this limit, advanced representations have been recently proposed. These intermediate languages, such as SignWriting [333] and the Hamburg Notation System (HamNoSys) [266], offer richer representations through visual and graphical symbols and proved to be promising as intermediaries for improving SLT. Leveraging such systems and external linguistic resources can address data scarcity and enhance the quality of sign language translations. We focus on the translation step and consider an ASCII encoding of SignWriting, namely Formal Sign-Writing (FSW), as a sign language representation. However our approach can be generalized to other representation languages.

Based on Bulla et al. [59], we present Advanced Use of LLMs for Sign Language Translation (AulSign), a novel sign language translation method that can handle languages not well represented in the LLM training data. Our method leverages on formalized lexicons specific to a given domain, incorporating external vocabularies to enhance translation. By employing few-

shot learning, our method significantly reduces the need for large training corpora, while introducing a novel, transparent, and explainable translation process at each stage. Our method, which addresses both spoken-to-sign and sign-to-spoken translation tasks, comprises three core components: a Retriever, an LLM and a Sign Mapper. The Retriever module identifies and retrieves samples from a training set which are used to instruct the LLM to map the input sentence into a pseudo-language that represent the sign language as a sequence of univocal descriptions of signs. The set of samples, enriched with formal grammatical rules, is integrated into the prompt to provide the LLM with a comprehensive linguistic and structural context. For spoken-to-sign, the LLM generates the corresponding translation in pseudo-language, which is then converted into the target language by the Sign Mapper, by mapping each part of the sequence to a predefined lexicon. The process is mirrored for the inverse sign-to-spoken translation task.

We evaluate our method on two datasets in two different spoken and sign languages: spoken Italian from and to LIS (Italian Sign Language) and spoken English from and to ASL (American Sign Language). Our experiments demonstrate the superiority of our method over state-of-the-art models for SignWriting translation, and show how LLM capabilities can be utilized for translation from and to unknown languages. Mor importantly, this advancement has the potential to enhance accessibility for the Deaf community.

The Chapter is structured as follows: Section 6.1 introduces the AulSign model, detailing its components and functionalities. Section 6.2 describes the experimental setup and presents results for both translation tasks in LIS and ASL. Section 6.3 analyzes the findings, Finally, section 6.4 presents the limitations of our work. Chapter 2.3 outlines the theoretical background for sign language tasks and reviews state-of-the-art methods for sign language

translation.

6.1 Methodology

Our method comprises three main components: a Retriever, an LLM, and a Sign Mapper. The Retriever identifies samples from the training set to be included in the prompt to instruct the LLM. The training set is pre-processed offline to convert signs into canonical descriptions, therefore each sample contains the spoken text and a corresponding decomposition into a sequence of canonical descriptions. The retrieved samples are used for prompt generation. The prompt generation combines the retrieved sample sentences, paired with their corresponding decompositions, with samples from a structured vocabulary and a set of predefined grammar rules to employ a structured prompt, enabling the LLM to perform few-shot inference on the input sentence. The Sign Mapper is employed only in spoken-to-sign translation and convert the sequence of canonical descriptions generated by the LLM into a sequence of signs. The whole process is described in Figure 6.1. Next we detail the various parts of our pipeline in the spoken-to-sign verse. The inverse process is similar, with small adjustments, which are discussed at the end of the section.

Our method considers a training set D of spoken sentences associated with their FSW counterpart, i.e. sequences of signs, and a vocabulary of signs V_t, where each sign is associated with one or more descriptions in natural language. We pre-process both D and V_t to translate FSW sequences into sequences of canonical descriptions[1]. First we define an equivalence operator $\equiv$[2] between signs and merge equivalent signs into one single entry,

[1] for Italian we do not need this step as in our dataset all signs have been manually associated to canonical descriptions

[2] in our implementation signs are considered equivalent if they contain the same set of

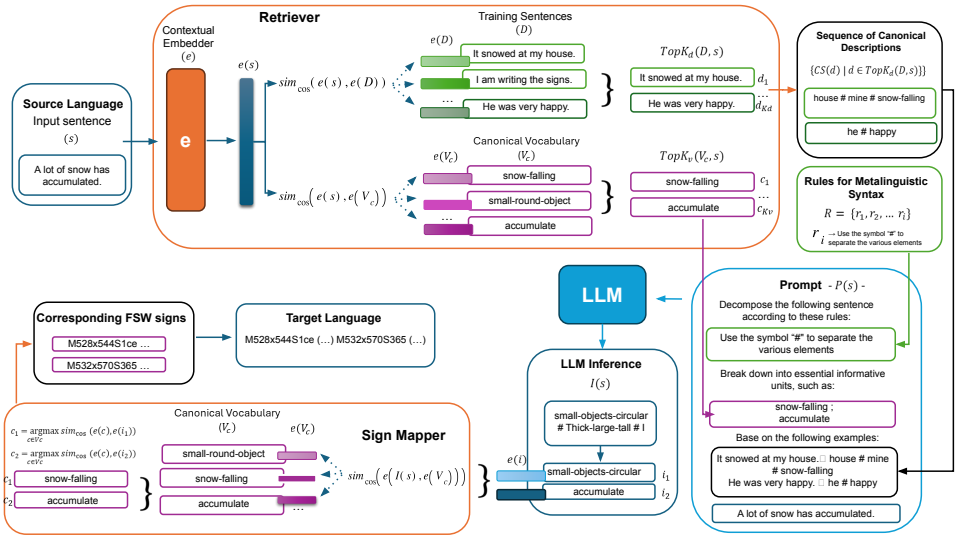


Figure 6.1: Overview of the spoken-to-sign AulSign pipeline.

choosing the most frequent sign as a representative. Then, we construct a set V_c of unique and non-ambiguous descriptions that we call *canonical descriptions*, by choosing among all descriptions associated to a sign the most frequent ones and combining them into a single string. Finally we substitute the sequences of signs associated to sentences in D with corresponding sequences of canonical descriptions by probing, for each sign of the sequence, an equivalent sign from V_t and taking the corresponding canonical description. If an equivalent sign exists in V_t, it is unique because all equivalent signs have been previously merged in V_t. On the other hand, if an equivalent sign is not found, a special "¡unk¿" description is considered. Given a spoken sentence $d \in D$ we denote as $CS(d)$ its associated sequence of canonical descriptions.

The Retriever identifies the top-K sentences from the training set (D) that are most semantically similar to the given input sentence (s). This process ensures that the subsequent prompt generation operates with con-

symbols with associated orientation and rotation; more complex notions that also consider the spatial position can be employed to improve results

textually relevant examples. To improve results, the Retriever also identifies a set of canonical descriptions from V_c that are similar to the input sentence, to feed to the LLM as samples of canonical descriptions. Both retrieval steps employ a contextual embedder $e(\cdot)$, trained for semantic similarity, to encode both the input sentence and the candidate sentences into high-dimensional representations. The similarity between s and each candidate is computed by cosine similarity. The system selects the top-K most semantically similar entries, which we denote with the sets $TopK_d(D, s)$ and $TopK_v(V_c, s)$, respectively.

The prompt generation step develops an input-sentence-customized prompt, leveraging on the few-shot learning abilities of LLMs. We design the prompt by combining grammar rules ($R = \{r_1, r_3, \ldots r_i\}$) with illustrative examples – retrieved sentences with their canonical descriptions – and the input sentence. The retrieved canonical descriptions from the Retriever are also included in the prompt as examples of canonical descriptions that the LLM can use to break down the input sentence. The complete prompt structure is shown in Figure 6.1.

The LLM generates a sequence $I(s)$ of strings in a form that mimic canonical descriptions in V_c, which feeds the Sign Mapper. To match the generated pseudo-canonical descriptions to entries in V_c we again use semantic similarity. We employ the same contextual embedder $e(\cdot)$ used for retrieval and, for each pseudo-canonical $i_l \in I(s)$, we extract: $c_l = \arg\max_{c \in V_c} \text{sim}_{\cos}(e(c), e(i_l))$. Finally, the model retrieves the corresponding FSW signs associated with the selected canonical descriptions to generate the final translation.

For the sign-to-spoken task, the input sentence in FSW is first converted into a sequence of canonical descriptions in the same way as for the train-

ing set pre-processing. Signs of the input FSW sequence are probed from V_t using the equivalence operator $\equiv$ and the corresponding canonical descriptions are taken. Similar sequences of canonical descriptions are then extracted from D and used with their speech counterpart to instruct the LLM, i.e. to build the prompt with few-shot samples. The prompt is built in the same way as for speech-to-sign, except that the examples for the break down are not given, since they are not necessary. The output of the LLM is then returned as the resulting spoken sentence.

6.2 Experiments and Results

To assess the effectiveness of AulSign, we conduct experiments on two benchmark datasets: the English SignBank+ [237] for ASL and an Italian dataset based on SignPuddle[3] for LIS. SignBank+ is a multilingual corpus with 254,002 distinct elements, 76 sign language encodings, and 153 "puddles", systems that categorize signs by language or dialect. The dataset is organized into three subcorpora: Manually (cleaned via manual review), GPT-3.5 (cleaned using GPT-3.5 with a few-shot learning paradigm [52]), and Bible (aligned biblical texts). Following preprocessing to ensure consistency and reduce noise, we focus exclusively on the English subset for our experiments. The English subcorpus of SignBank+ comprises 43,705 annotated items spanning nine variants of English Sign Language, including 19,304 unique signs and 13,631 sentences. We use the unique-sign English SignBank+ subcorpus to build V_c (cf. Sect. 6.1) and split the sentence-based subcorpus into training set D (cf. Sect. 6.1) and test set. We experiment with three different training set configurations (referred to as AulSign I, AulSign II, AulSign III) to assess the method's robustness across varying

[3]https://www.signbank.org/signpuddle/

data conditions. AulSign I considers the entire training set, which counts 13,275 sentences. Note that not all signs in the sentences are covered by our vocabulary (extracted by single sign sentences), therefore our method, which requires a complete vocabulary, is at a disadvantage in this setting[4]. AulSign II filters out sentences containing out-of-vocabulary signs, resulting in a more consistent dataset of 2,301 sentences. AulSign III simulates a low-resource scenario by randomly sampling 115 sentences from the AulSign II training set. For comparisons, we fine-tuned the state-of-the-art model from Jiang et al. [164] using three parallel configurations, denoted as Jiang et al. I, II, and III. Each model is trained on the same set of training sentences as the corresponding AulSign configurations and includes the complete vocabulary in the training process. Following Jiang et al., we configure the set of hyperparameters with a learning rate of 0.0001, a drop-out of 0.50, a batch size of 64[5] and a learning-rate-factor of 0.70. We evaluate both approaches on a test set of 356 sentences.

The Italian SignPuddle corpus contains 2,149 annotated items, including 1,974 signs and 782 sentences. Each sign is associated with a hand-made canonical description and a FSW representation. We use signs to populate the vocabulary V_c and split the dataset into training and test sets of 547 and 235 sentences, respectively.

To retrieve relevant examples for the AulSign Retriever module, we employ the Lee et al. [198] and the Reimers et al. [279] models for ASL and LIS, respectively, as embedding models. We set the number of top retrieved sentences K_d to 20 and the number retrieved canonical descriptions K_v to 100.

[4]we identify the extraction of vocabulary from sentences as an interesting research direction

[5]in place of 32 used by Jiang et al. to improve efficiency

For prompt generation (cf. Sect. 6.1), we define two distinct sets of rules to structure the metalinguistic syntax settings. These rules describe each sign based on its canonical description and are supplemented with illustrative examples of sentences and their associated canonical descriptions to ensure clarity and alignment between linguistic concepts and their representations. We define a total of seven rules for both English and Italian[6]. For all experiments, we employ the GPT-3.5 as LLM.[7]

Section 6.2.1 details the overall performance of the AulSign method for LIS and ASL in executing the spoken-to-sign translation task. Section 6.2.2 presents the results for the inverse task, i.e., sign-to-spoken translation.

6.2.1 Spoken-to-sign translation

Table 6.1 presents the overall performance of the AulSign method on the English SignBank+ dataset in terms of F1-score, Bleu, ChrF2, and Mean Absolute Error (MAE). We consider the F1-score as an appropriate metric to evaluate predictions at the symbol (with associated rotation and orientation) level, since the order of symbols within a sign is not semantically relevant. However, the order across signs is semantically relevant, therefore symbol-level evaluation alone is insufficient for FSW. To address this limitation, and following Jiang et al. [164], we also employ the popular translation metrics BLEU [257] and ChrF2 [264], which capture statistics at both word and character levels. Additionally, we assess positional values (x and y) using MAE to quantify the discrepancy between predicted and ground-truth values. As the symbol order within a sign is not semantically significant, both gold standard and predicted sequences are alphabetically sorted before

[6]for an overview of the grammar rules employed, we refer to https://anonymous.4open.science/r/AulSign-13F6

[7]https://platform.openai.com/docs/models/gpt-3-5-turbo

evaluation.

As shown in Table 6.1, AulSign consistently outperforms the Jiang et al. baseline in data-scarce scenarios. In the low-resource configuration (AulSign III), the model achieves an F1 score of 0.38, an 11-point improvement over Jiang et al. III (i.e. 0.27). This trend is mirrored in BLEU and ChrF2 scores, where AulSign III attains 20.40 and 54.14, compared to 10.94 and 39.17 for the baseline. These results underscore AulSign's robustness in low-resource settings. In data-rich settings (AulSign I and II), AulSign demonstrates comparable performance to Jiang et al., even when confronted with noisy data. AulSign I achieves an F1 score of 0.42, slightly lower than Jiang et al. I (i.e. 0.45), while maintaining competitive BLEU (i.e. 25.40 vs. 29.26) and ChrF2 scores (i.e. 56.44 vs. 57.82). Our assessment affirms AulSign's robustness across different data conditions. Regarding MAE, Jiang et al.'s models generally achieve lower positional errors, with minimum values of 23.78 and 27.80 for X and Y coordinates, respectively. AulSign's MAE values are slightly higher (e.g., 25.02 and 29.66 for AulSign I), reflecting a trade-off between precise factor prediction and overall symbol recognition quality. This discrepancy can be attributed to the architectural focus of Jiang et al.'s models on positional accuracy, whereas AulSign employs a general-purpose translation framework. Nevertheless, AulSign achieves competitive overall performance, highlighting its efficacy in handling both sequence-level and positional challenges in SignWriting-based translation. To better illustrate the performance trend as the amount of training data varies, we report in Figure 6.2 the F1-score column of Table 6.1. AulSign shows consistent performance across training set sizes, contrasting with Jiang et al.'s model, which exhibits significant degradation in low-resource conditions.

Table 6.2 summarizes the performance of AulSign on the Italian Sign-

Model	F1	BLEU	ChrF2	MAE X	MAE Y	Training size
Jiang et al. I	**0.45**	**29.26**	**57.82**	**23.77**	**27.96**	13,275
AulSign I	0.42	25.40	56.44	25.02	29.66	
Jiang et al. II	0.40	22.68	50.65	**23.78**	**27.80**	2,301
AulSign II	**0.41**	**23.96**	**56.06**	25.42	30.31	
Jiang et al. III	0.27	10.94	39.17	**23.81**	**27.82**	115
AulSign III	**0.38**	**20.40**	**54.14**	25.50	30.55	

Table 6.1: Overall spoken-to-sign ASL translation models performance on the English SignBank+ dataset, in terms of F1 score, BLEU, ChrF2, and MAE. We test three AulSign configurations: (I) data-rich (13,275 sentences), (II) filtered (2,301 sentences), and (III) low-resource (115 sentences). We compare the results to the baseline model (Jiang et al. [164]) under the same conditions, with bold values indicating the best result per configuration.

Model	F1	BLEU	ChrF2	MAE X	MAE Y
Jiang et al.	0.50	16.40	45.18	23.82	37.44
AulSign	**0.63**	**37.71**	**57.54**	**21.22**	**33.22**

Table 6.2: Models overall performance in spoken-to-sign Italian-to-LIS translation using FSW-encoded sequences on the Italian SignPuddle corpus.

Puddle corpus, evaluated using F1-score, BLEU, ChrF2, and MAE for positional coordinates. AulSign achieves substantial improvements across all metrics compared to Jiang et al. Notably, it attains an F1 score of 0.63, representing a 13-point increase over the baseline (0.50), demonstrating superior capability in sign classification and identification. For sequence-level metrics, AulSign outperforms Jiang et al. with BLEU and ChrF2 scores of 37.71 and 57.54, respectively, compared to the baseline's 16.40 and 45.18. Additionally, AulSign demonstrates superior positional accuracy, with MAE values of 21.22 (X) and 33.22 (Y), compared to 23.82 and 37.44 for Jiang et al. This indicates AulSign's ability to precisely model even the spatial characteristics of signs when a comprehensive vocabulary is available.

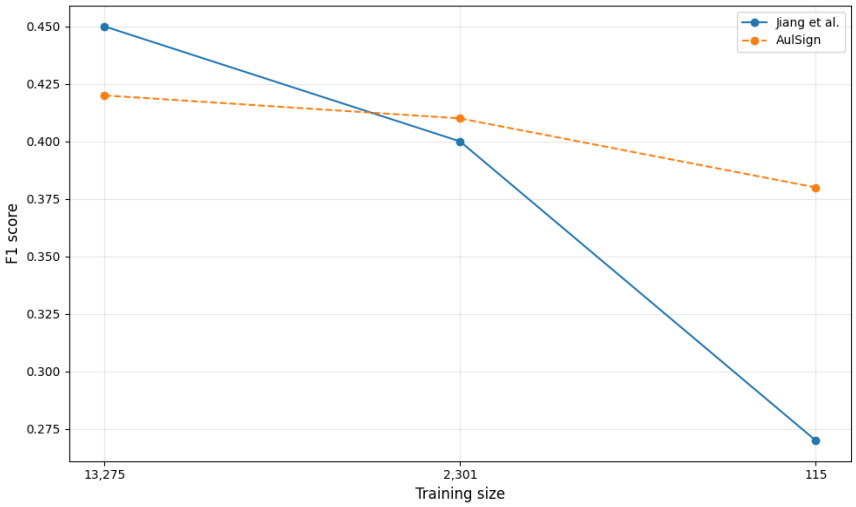


Figure 6.2: Comparative analysis between AulSign and the state-of-the-art model across different training set sizes, evaluated in terms of F1 score for the ASL spoken-to-sign translation task.

6.2.2 Sign-to-spoken translation

Tables 6.3 and 6.4 show the results on sign-to-spoken translation for ASL-to-English and LIS-to-Italian, using FSW-encoded sign language sequences. As for spoken-to-sign, we evaluate three AulSign model configurations (I, II, III) for ASL and one for LIS in terms of BLEU and ChrF2. We do not compute F1 and MAE since the former is not adequate to assess spoken natural language sentences and the latter has no meaning in this context. For a comprehensive overview of model configurations and experimental setup, we refer to the beginning of this section.

As shown in Table 6.3 and Table 6.4, AulSign consistently outperforms the baseline, achieving state-of-the-art results for both ASL and LIS. In a low-resource setting for the ASL, AulSign significantly outperforms the Jiang et al. model, yielding a BLEU score of 26.59 and a ChrF2 score of 40.76, compared to the baseline's 2.20 and 18.90, respectively. Similarly, AulSign achieves a BLEU score of 17.95 and a ChrF2 score of 50.42, significantly

Model	BLEU	ChrF2	Training Size
Jiang et al. I	18.40	34.40	13,275
AulSign I	**23.38**	**39.21**	
Jiang et al. II	8.40	22.50	2,301
AulSign II	**24.75**	**40.26**	
Jiang et al. III	2.20	18.90	115
AulSign III	**26.59**	**40.76**	

Table 6.3: Overall sign-to-spoken ASL translation models performance on the English SignBank+ dataset, in terms of BLEU, and ChrF2. We test three AulSign configurations: (I) data-rich (13,275 sentences), (II) filtered (2,301 sentences), and (III) low-resource (115 sentences). We compare the results to the baseline model (Jiang et al. [164]) under the same conditions. Bold values indicate the best result per configuration.

Model	BLEU	ChrF2
Jiang et al.	6.50	29.40
AulSign	**17.95**	**50.42**

Table 6.4: Models overall performance in sign-to-spoken LIS translation using FSW-encoded sequences on the Italian SignPuddle dataset, evaluated in terms of BLEU and ChrF2.

outperforming the baseline scores of 6.50 and 29.40 in the low-resource LIS setting. In high-resource settings, we evaluate two configurations of AulSign: AulSign I and AulSign II. Both configurations show superior performance compared to the Jiang et al. model. Specifically, AulSign I achieves a BLEU score of 23.38 and a ChrF2 score of 39.21, exceeding the baseline scores of 18.40 and 34.40. Notably, AulSign II, trained on a reduced dataset, significantly outperforms the baseline, with a BLEU score of 24.75 and a ChrF2 score of 40.26, compared to the baseline's 8.40 and 22.50. In low-resource scenarios the low performances of Jiang et al. are expected since there are not sufficient data for learning, considering the complexity of natural language. In any case, our model can leverage on the language processing abilities of pretrained LLMs, therefore producing higher-quality translation.

6.3 Discussion

This study presents the evaluation of AulSign for SLT, particularly in low-resource scenarios, using SignWriting as an intermediate representation. AulSign achieves state-of-the-art results for both ASL and LIS in spoken-to-sign and sign-to-spoken tasks, consistently outperforming Jiang et al. [164] across different data configurations. Notably, AulSign excels in data-scarce environments, demonstrating strong generalization and maintaining performance even with reduced training data.

AulSign shows to effectively produce linguistically accurate and spatially coherent sign language translations, even with limited data. Its strong performance in sign-to-spoken tasks, reflected in high BLEU and chrF2 scores, is due to its use of context augmentation and domain-specific vocabulary mapping. AulSign adapts well to varying data sizes, maintaining consistent results, unlike the Jiang et al. model, which struggles with smaller datasets. While AulSign is slightly behind the baseline in some data-rich scenarios, its adaptability is superior, especially with smaller or higher-quality datasets.

Across tasks, AulSign demonstrates balanced performance. For ASL, the model achieves average BLEU scores of 23.25 in spoken-to-sign and 24.90 in sign-to-spoken, with chrF2 scores slightly higher in the latter (55.54 vs. 40.07). This disparity is likely due to the structured prompt in sign-to-spoken, which provides explicit linguistic context for inference. In LIS translation, AulSign consistently outperforms Jiang et al., benefiting from the highest quality level of the Italian SignPuddle dataset. AulSign's approach leverages on a predefined, domain-specific vocabulary, which plays a central role in improving token alignment and translation accuracy. This strategy ensures precise mappings between signs and their representations,

particularly in structured intermediate formats like SignWriting. However, signs not present in the vocabulary are treated as unknown, which may affect the model's ability to handle dynamic linguistic contexts or rare signs. Expanding the vocabulary to cover a broader range of linguistic variations and incorporating mechanisms for inferring or adapting to unseen signs would further enhance the model's capacity to address diverse translation challenges. This is particularly relevant in high-resource settings, where the model's performance is highly sensitive to vocabulary size and coverage. Finally, AulSign's modular architecture ensures transparency at each stage of the translation pipeline. The retrieval module, for instance, selectively augments training samples, while the prompt generation phase explicitly aligns input sentences with grammatical rules. This design allows for targeted error analysis, such as identifying misalignments between FSW symbols and canonical descriptions, which is infeasible in end-to-end SLT models. By prioritizing explainability, AulSign bridges the gap between high performance and user trust, offering a reliable tool for researchers and end-users.

6.4 Limitations

While AulSign demonstrates promising results, certain limitations affect its generalizability and broader applicability. The model relies on a vocabulary of canonical descriptions to map signs to natural language, ensuring consistency and control over translation quality. However, this dependency introduces a degree of rigidity, as it relies on a specialized vocabulary in which a one-to-one correspondence between signs and their descriptions must be explicitly defined. Furthermore, although AulSign is designed to perform effectively in low-resource settings, its translation quality remains contingent on the availability and consistency of training data. Incomplete or incon-

sistently annotated datasets may adversely impact accuracy, particularly in capturing the spatial positioning of signs. At present, this study employs SignWriting as the primary intermediate representation due to its structured and expressive nature. Nonetheless, the approach is inherently extensible to alternative notational systems, such as HamNoSys or specialized lexical glossaries, thereby offering potential avenues for enhanced adaptability and integration within automatic sign language translation frameworks.

7 Conclusion

This dissertation examines the integration of HCAI principles into NLP, addressing the interrelated challenges of moral value alignment, controllability, evaluation of subjective tasks, and inclusivity for underrepresented communities. Across theoretical, methodological, and applied dimensions, the contributions collectively advance the design of NLP systems that are more aligned with human values, transparent in their behavior, and accessible to different users.

We demonstrate the potential of LLMs for moral value detection, with performance shaped by model size, prompting strategies, and domain variation [53]. While GPT-4 achieves the strongest results, smaller models such as Mistral also show promise in resource-constrained settings. These findings introduce prompting strategies for applying current models to moral value detection grounded in MFT, achieving human-comparable performance and advancing the integration of such models into ethical AI systems.

To strengthen evaluation in subjective and ethically sensitive domains, we introduce F1-kappa, a novel metric that combines precision–recall dynamics with chance-corrected agreement, enabling more robust and human-centered assessment of model performance [58]. By capturing the variability of human annotations, F1-kappa improves accountability and fairness in evaluation.

We further present GRAMMAR-LLM, a framework that integrates formal grammatical constraints into language model decoding, enforcing syntactic compliance in real time with minimal computational cost [339]. GRAMMAR-LLM improves classification accuracy, translation quality, and structural fidelity across tasks such as hierarchical classification, sign language translation, and semantic parsing. By constraining generation, this approach advances the development of AI systems that are both expressive and human-controllable.

Finally, we introduce AulSign, a framework for SLT that combines retrieval-augmented generation with SignWriting as an intermediate representation [59]. AulSign achieves state-of-the-art performance in spoken-to-sign and sign-to-spoken translation for English–ASL and Italian–LIS, excelling in low-resource scenarios. Through its structured vocabularies and modular, explainable architecture, it enhances translation accuracy, interpretability, and error analysis. This contribution represents a step toward more inclusive AI, particularly for the Deaf community.

Overall, this dissertation advances beyond a performance-centric paradigm of AI by promoting systems that are responsible, interpretable, and genuinely human-centered. Through the integrated pursuit of ethical alignment, methodological rigor, and inclusivity, it establishes a coherent framework for the development of NLP technologies that not only achieve technical excellence but also contribute to human well-being and the broader societal good.

Bibliography

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

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] EU Artificial Intelligence Act. The eu artificial intelligence act. *European Union*, 2024.

[4] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24: 1750–1762, 2021.

[5] Kareem Ahmed, Kai-Wei Chang, and Guy Van den Broeck. Controllable generation via locally constrained resampling. *arXiv preprint arXiv:2410.13111*, 2024.

[6] Alfred V Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques, and Tools*. Pearson Education, Inc, 2006.

[7] Hleg Ai. High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, 6, 2019.

[8] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[9] Ramya Akula and Ivan Garibay. Ethical ai for social good. In *International Conference on Human-Computer Interaction*, pages 369–380. Springer, 2021.

[10] Mark Alfano, Marc Cheong, and Oliver Scott Curry. Moral universals: A machine-reading analysis of 256 societies. *Heliyon*, 10(6), 2024.

[11] Ameera M Almasoud and Hend S Al-Khalifa. A proposed semantic machine translation system for translating arabic text to arabic sign language. In *Proceedings of the Second Kuwait Conference on e-Services and e-Systems*, pages 1–6, 2011.

[12] Ameera M Almasoud and Hend S Al-Khalifa. Semsignwriting: A proposed semantic system for arabic text-to-signwriting translation. 2012.

[13] Milad Alshomary and et al. Moralarg: A large-scale argumentation dataset with moral foundations. In *Proceedings of ...*, 2022.

[14] Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. The moral debater: A study on the computational gener-

ation of morally framed arguments. *arXiv preprint arXiv:2203.14563*, 2022.

[15] Steven Alter. Agent responsibility framework for digital agents: roles and responsibilities related to facets of work. In *International Conference on Business Process Modeling, Development and Support*, pages 237–252. Springer, 2022.

[16] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.

[17] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[18] Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. Exploiting biased models to de-bias text: A gender-fair rewriting model. *arXiv preprint arXiv:2305.11140*, 2023.

[19] Krasimir Angelov. Incremental parsing with parallel multiple context-free grammars. In *Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009)*, pages 69–76, 2009.

[20] Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284, 2022.

[21] Obinna Anya and Hissam Tawfik. An "awareness" environment for clinical decision support in e-health. In *International Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018)*, pages 456–467, Granada, Spain, 2018. Springer. doi: 10.1007/978-3-319-78759-6_41.

[22] Oscar Araque and et al. Moralstrength: A moral lexicon for moral values detection. 2020.

[23] Oscar Araque and et al. Libertymfd: Extending moral lexicons with liberty values. 2022.

[24] Luigi Asprino, Stefano De Giorgis, Aldo Gangemi, Luana Bulla, Ludovica Marinucci, Misael Mongiovi, et al. Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 33–41. Association for Computational Linguistics, 2022.

[25] Mohammad Atari and et al. The moral foundations questionnaire–2 (mfq-2). 2023.

[26] Mohammad Atari and Jonathan Haidt. Ownership is (likely to be) a moral foundation. *Behavioral & Brain Sciences*, 46, 2023.

[27] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 2023.

[28] Mohammad Atari, Matthias R Mehl, Jesse Graham, John M Doris, Norbert Schwarz, Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Elaine Gonzalez, Nikki Jafarzadeh, et al. The paucity of morality in everyday talk. *Scientific Reports*, 13(1):5967, 2023.

[29] Jan Auernhammer. Human-centered ai: The role of human-centered design research in the development of ai. 2020.

[30] Tina Babu, Rekha R Nair, et al. Emotion-aware music recommendation system: Enhancing user experience through real-time emotional context. *arXiv preprint arXiv:2311.10796*, 2023.

[31] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PloS one*, volume 10, page e0130140. Public Library of Science, 2015.

[32] Xuefeng Bai, Yulong Chen, and Yue Zhang. Graph pre-training for amr parsing and generation. *arXiv preprint arXiv:2203.07836*, 2022.

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

[34] Jason Bailey, Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, et al. Participatory approaches to machine learning. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and*

Transparency, pages 1–12, 2022. doi: 10.1145/3531146.3533152. URL https://dl.acm.org/doi/10.1145/3531146.3533152.

[35] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013.

[36] Mariano Beiró and et al. Moral values in covid-19 vaccine discourse: Facebook comments dataset. 2023.

[37] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[38] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, New York, NY, USA, 2021. ACM. doi: 10.1145/3442188.3445922.

[39] Jürgen Bernard, Marco Hutter, Heiko Reinemuth, Hendrik Pfeifer, Christian Bors, and Jörn Kohlhammer. Visual-interactive preprocessing of multivariate time series data. *Computer Graphics Forum*, 38 (3):401–412, 2019. doi: 10.1111/cgf.13698.

[40] Steve J Bickley and Benno Torgler. Cognitive architectures for artificial intelligence ethics. *Ai & Society*, 38(2):501–519, 2023.

[41] Pablo Biedma, Xiaoyuan Yi, Linus Huang, Maosong Sun, and Xing Xie. Beyond human norms: Unveiling unique values of large lan-

guage models through interdisciplinary approaches. *arXiv preprint arXiv:2404.12744*, 2024.

[42] Ahsan Bilal, David Ebert, and Beiyu Lin. Llms for explainable ai: A comprehensive survey. *arXiv preprint arXiv:2504.00125*, 2025.

[43] Sergiu Bilc, Adrian Groza, George Muntean, and Simona Delia Nicoara. Interleaving automatic segmentation and expert opinion for retinal conditions. *Diagnostics*, 12(1):22, 2021. doi: 10.3390/diagnostics12010022.

[44] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, pages 149–159. ACM, 2018. doi: 10.1145/3287560.3287582. URL `https://dl.acm.org/doi/10.1145/3287560.3287582`.

[45] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, 2021.

[46] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.485.

[47] Julia Bonn, Skatje Myers, Jens Van Gysel, Lukas Denk, Meagan Vigus, Jin Zhao, Andrew Cowell, William Croft, Jan Hajic, Martha Palmer, et al. Abstract meaning representation (amr) annotation release 3.0.

[48] Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. Bias and fairness in mul-

timodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 268–277, 2021.

[49] Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. Investigating human values in online communities. *arXiv preprint arXiv:2402.14177*, 2024.

[50] Simone Borsci, Ville V Lehtola, Francesco Nex, Michael Ying Yang, Ellen-Wien Augustijn, Leila Bagheriye, Christoph Brune, Ourania Kounadi, Jamy Li, Joao Moreira, et al. Embedding artificial intelligence in society: looking beyond the eu ai master plan using the culture cycle. *AI & society*, 38(4):1465–1484, 2023.

[51] Karen L Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

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

[53] Luana Bulla, Misael Mongiovì, Stefano De Giorgis, and Aldo Gangemi. Large language models meet moral values: A comprehensive assessment of moral abilities. *Available at SSRN 4907562*.

[54] Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Chiara Lucifora, and Misael Mongiovì. Comparing user perspectives in a virtual reality cul-

tural heritage environment. In *International conference on advanced information systems engineering*, pages 3–15. Springer, 2023.

[55] Luana Bulla, Aldo Gangemi, and Misael Mongiovì. Do language models understand morality? towards a robust detection of moral content. In *International Workshop on Value Engineering in AI*, pages 98–113. Springer, 2023.

[56] Luana Bulla, Aldo Gangemi, et al. Towards distribution-shift robust text classification of emotional content. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8256–8268, 2023.

[57] Luana Bulla, Stefano De Giorgis, Misael Mongiovì, and Aldo Gangemi. Large language models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports*, 17:100609, 2025. ISSN 2451-9588. doi: https://doi.org/10.1016/j.chbr.2025.100609. URL `https://www.sciencedirect.com/science/article/pii/S2451958825000247`.

[58] Luana Bulla, Misael Mongiovì, and Aldo Gangemi. Underperformance or pluralism: A machine learning perspective on inter-annotator agreement. In *Proceedings of the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI 2025)*, Pisa, Italy, June 2025. IOS Press. URL `https://hhai-conference.org/2025/`.

[59] Luana Bulla, Gabriele Tuccio, Misael Mongiovì, and Aldo Gangemi. Leveraging large language models for accurate sign language translation in low-resource scenarios. *arXiv preprint arXiv:2508.18183*, 2025.

[60] P Burggräf, J Wagner, and TM Saßmannshausen. Sustainable interaction of human and artificial intelligence in cyber production man-

agement systems. In *Congress of the German Academic Association for Production Technology*, pages 508–517. Springer, 2020.

[61] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1): 2053951715622512, 2016.

[62] Shu Cai and Kevin Knight. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, 2013.

[63] Scott Allen Cambo and Darren Gergle. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[64] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549, 2018.

[65] Tara Capel and Margot Brereton. What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–23, 2023.

[66] Carlos Carrasco-Farre. Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of llm arguments. *arXiv preprint arXiv:2404.09329*, 2024.

[67] Margarita Robles Carrillo. Artificial intelligence: From ethics to law. *Telecommunications policy*, 44(6):101937, 2020.

[68] Andreas Cebulla, Zygmunt Szpak, Catherine Howell, Genevieve Knight, and Sazzad Hussain. Applying ethics to ai in the workplace: the design of a scorecard for australian workplace health and safety. *AI & society*, 38(2):919–935, 2023.

[69] Joana Cerejo and Miguel Carvalhais. The lens of anticipatory design under ai-driven services. In *International Conference on Digital Design & Communication (DIGICOM 2020) ATAS: PT & ES, Barcelos, Portugal*, pages 345–357, 2020.

[70] Rémy Chaput, Jérémy Duval, Olivier Boissier, Mathieu Guillermin, and Salima Hassas. A multi-agent approach to combine reasoning and learning for an ethical behavior. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 13–23, 2021.

[71] Khansa Chemnad and Achraf Othman. Digital accessibility in the era of artificial intelligence—bibliometric analysis and systematic review. *Frontiers in Artificial Intelligence*, 7:1349668, 2024.

[72] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[73] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Gen-

erating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[74] Scott Clifford and et al. Moral foundations vignettes: A standard set of moral scenarios. In *Proceedings of ...*, 2015.

[75] Mike Cooley. Human-centered design. *Information design*, pages 59–81, 2000.

[76] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, 2022.

[77] Rita Cucchiara and Matteo Fabbri. Fine-grained human analysis under occlusions and perspective constraints in multimedia surveillance. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–23, 2022.

[78] Oliver Scott Curry. Morality as cooperation: A problem-centred approach. In *The evolution of morality*, pages 27–51. Springer, 2016.

[79] Oliver Scott Curry, Mark Alfano, Mark J Brandt, and Christine Pelican. Moral molecules: Morality as a combinatorial system. *Review of Philosophy and Psychology*, 13(4):1039–1058, 2022.

[80] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[81] Devleena Das and Sonia Chernova. Leveraging rationales to improve human task performance. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 510–518, 2020.

[82] Subhajit Das, Dylan Cashman, Remco Chang, and Alex Endert. Gaggle: Visual analytics for model space navigation. In *Proceedings of Graphics Interface (GI '20)*, pages 137–147. University of Toronto, 2020. doi: 10.20380/GI2020.15.

[83] Mehdi Dastani and Vahid Yazdanpanah. Responsibility of ai systems. *Ai & Society*, 38(2):843–852, 2023.

[84] Fernando de Almeida Freitas, Sarajane Marques Peres, Otávio de Paula Albuquerque, and Marcelo Fantinato. Leveraging sign language processing with formal signwriting and deep learning architectures. In *Brazilian Conference on Intelligent Systems*, pages 299–314. Springer, 2023.

[85] Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23(3):1305–1331, 2024.

[86] Nour El Houda Dehimi and Zakaria Tolba. Attention mechanisms in deep learning: Towards explainable artificial intelligence. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–7. IEEE, 2024.

[87] Leon Derczynski. Complementarity, f-score, and nlp evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, 2016.

[88] Daniel Deutsch, Shyam Upadhyay, and Dan Roth. A general-purpose algorithm for constrained sequential inference. In *Proceedings of*

the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 482–492, 2019.

[89] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[90] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*, 2023.

[91] Frank Dignum and Virginia Dignum. How to center ai on humans. In *NeHuAI 2020, 1st International Workshop on New Foundations for Human-Centered AI, Santiago de Compostella, Spain, September 4, 2020*, pages 59–62, 2020.

[92] Murat Dikmen and Catherine Burns. The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162: 102792, 2022.

[93] Yixin Dong, Charlie F Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. Xgrammar: Flexible and efficient structured generation engine for large language models. *arXiv preprint arXiv:2411.15100*, 2024.

[94] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. Trust in automl: exploring information needs for establishing trust in automated

machine learning systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 297–307, 2020.

[95] Shitong Duan, Xiaoyuan Yi, Peng Zhang, Dongkuan Xu, Jing Yao, Tun Lu, Ning Gu, and Xing Xie. Adaem: An adaptively and automated extensible measurement of llms' value difference. *arXiv preprint arXiv:2505.13531*, 2025.

[96] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[97] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*, 2018.

[98] Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. Empirical methodology for crowdsourcing ground truth. *Semantic Web*, 12(3):403–421, 2021.

[99] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012. doi: 10.1145/2090236.2090255. URL `https://dl.acm.org/doi/10.1145/2090236.2090255`.

[100] Christos Emmanouilidis, Sabine Waschull, Jos Bokhorst, and Hans Wortmann. Human in the ai loop in production environments. In *IFIP International Conference on Advances in Production Manage-*

ment Systems (APMS '21), pages 331–342, Cham, 2021. Springer. doi: 10.1007/978-3-030-85910-7_35.

[101] Daniel Estrada. Human supremacy as posthuman risk. *The Journal of Sociotechnical Critique*, 1(1):5, 2020.

[102] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). COM(2021) 206 final, 2021. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`.

[103] Gregory Ewing and Ibrahim Demir. An ethical decision-making framework with serious gaming: a smart water case study on flooding. *Journal of Hydroinformatics*, 23(3):466–482, 2021.

[104] Rosa Falotico and Piero Quatto. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470, 2015.

[105] Uzma Farooq, Mohd Shafry Mohd Rahim, Nabeel Sabir, Amir Hussain, and Adnan Abid. Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, 33(21):14357–14399, 2021.

[106] Pooya Fayyazsanavi, Antonios Anastasopoulos, and Jana Košecká. Gloss2text: Sign language gloss translation using llms and semantically aware label smoothing. *arXiv preprint arXiv:2407.01394*, 2024.

[107] Qi Feng, Debiao He, Zhe Liu, Huaqun Wang, and Kim-Kwang Raymond Choo. Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, 15:3709–3721, 2020.

[108] Ronald Fischer, Markus Luczak-Roesch, and Johannes A Karl. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*, 2023.

[109] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. Technical report, Berkman Klein Center for Internet & Society, 2020.

[110] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[111] Luciano Floridi. *The Ethics of Artificial Intelligence*. Oxford University Press, 2023.

[112] Francesca Foffano, Teresa Scantamburlo, and Atia Cortés. Investing in ai for social good: an analysis of european national strategies. *AI & society*, 38(2):479–500, 2023.

[113] Max Forbes and et al. Social chemistry 101: Norms for everyday situations. 2020.

[114] Jeremy Frimer. Moral foundations dictionary version 2. 2019.

[115] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[116] Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. Factored neural machine translation architectures. In *Proceedings of the 13th International Conference on Spoken Language Translation*, 2016.

[117] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M. Fiore, Ivan Garibay, Keri Grieman, John C. Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwowski, Joe Kider, Joseph Konstan, Sean Koon, Monica Lopez-Gonzalez, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ten Holter, and Wei Xu. Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39(3): 391–437, 2023. doi: 10.1080/10447318.2022.2153320. URL `https://doi.org/10.1080/10447318.2022.2153320`.

[118] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.

[119] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[120] Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured nlp tasks without finetuning. *arXiv preprint arXiv:2305.13971*, 2023.

[121] Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. Generating structured outputs from language models: Benchmark and studies. *arXiv preprint arXiv:2501.10868*, 2025.

[122] Lorenzo Genta et al. Consensus-based crowdsourcing: Techniques and applications. 2015.

[123] Niloofar Golazizian, Malihe Alikhani, and Snigdha Chaturvedi. Moral foundations subjective corpus: A framework for capturing annotator-specific moral judgments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[124] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372, 2024.

[125] Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, 55(14s):1–39, 2023.

[126] Jesse Graham and et al. The moral foundations dictionary. 2009.

[127] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.

[128] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.

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

[130] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[131] Siyi Guo, Negar Mokhberian, and Kristina Lerman. 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, 2023.

[132] Sai Abishek Gurrapu, Vivek Gupta, Srijan Kumar, Anshuman Shukla, et al. Rationalization for explainable nlp: A survey. *arXiv preprint arXiv:2301.10568*, 2023.

[133] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, 2020.

[134] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

[135] Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*, 2022.

[136] Farsheed Haque, Depeng Xu, and Xi Niu. A comprehensive survey on bias and fairness in large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 83–101. Springer, 2025.

[137] Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. Modular and on-demand bias mitigation

with attribute-removal subnetworks. *arXiv preprint arXiv:2205.15171*, 2022.

[138] Hongmei He, John Gray, Angelo Cangelosi, Qinggang Meng, T Martin McGinnity, and Jörn Mehnen. The challenges and opportunities of human-centered ai for trustworthy robots and autonomous systems. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4): 1398–1412, 2021.

[139] Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. Whose emotions and moral sentiments do language models reflect? *arXiv preprint arXiv:2402.11114*, 2024.

[140] Marco Hellmann, Diana C Hernandez-Bocanegra, and Jürgen Ziegler. Development of an instrument for measuring users' perception of transparency in recommender systems. *system*, 12:7, 2022.

[141] Thomas Herrmann and Sabine Pfeifer. Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. *AI & Society*, 2022. doi: 10.1007/s00146-022-01391-5.

[142] Hendrik Heuer and Andreas Breiter. How fake news affect trust in the output of a machine learning system for news curation. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pages 18–36. Springer, 2020.

[143] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*, pages 95–99, New York, NY, USA, 2017. ACM. doi: 10.1145/3064663.3064703.

[144] Katja Hofmann, Jack Clark, Kyunghyun Cho, and Sebastian Risi. Ai and computational creativity workshop. In *NeurIPS 2019 Workshops*, 201.

[145] Andreas Holzinger. The next frontier: Ai we can really trust. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 427–440. Springer, 2021.

[146] Andreas Holzinger, Michaela Kargl, Bettina Kipperer, Peter Regitnig, Markus Plass, and Heimo Müller. Personas for artificial intelligence (ai) an open source toolbox. *IEEE Access*, 10:23732–23747, 2022.

[147] Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Carl Orge Retzlaff, Andreas Gronauer, Vladimir Pejakovic, Francisco Medel-Jimenez, Theresa Krexner, Christoph Gollob, and Karl Stampfer. Digital transformation in smart farm and forest operations needs human-centered ai: challenges and future directions. *Sensors*, 22(8):3043, 2022.

[148] Joe Hoover, Gabriella Portillo-Wightman, Liane Yeh, and et al. Moral foundations twitter corpus: A dataset of moral values from social media. 2020.

[149] John E Hopcroft, Rajeew Motwani, and Jeffrey D Ullman. *Introduction to automata theory, languages and computation*. Addison-Wesley, 2001.

[150] Franziska Hopp and et al. Extended moral foundations dictionary (emfd). 2021.

[151] Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. The extended moral foundations dictionary (emfd): De-

velopment and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53(1):232–246, 2021.

[152] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. Dramatvis personae: Visual text analytics for identifying social biases in creative writing. In *Proceedings of the 2022 ACM designing interactive systems conference*, pages 1260–1276, 2022.

[153] Dirk Hovy, Shrimai Prabhumoye, et al. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8): e12432, 2021. doi: 10.1111/lnc3.12432.

[154] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005.

[155] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023.

[156] Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao. An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, 2022.

[157] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Prin-

ciples, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[158] Xiaolei Huang, Alexandra Wormley, and Adam Cohen. Learning to adapt domain shifts of moral values via instance weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 121–131, 2022.

[159] Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C Park. An efficient sign language translation using spatial configuration and motion dynamics with llms. *arXiv preprint arXiv:2408.10593*, 2024.

[160] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

[161] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[162] Julie Jiang and Emilio Ferrara. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893*, 2023.

[163] Julie Jiang, Luca Luceri, and Emilio Ferrara. Moral values underpinning covid-19 online communication patterns. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2642–2650, 2025.

[164] Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. Machine translation between spoken languages and signed languages represented in signwriting. *arXiv preprint arXiv:2210.05404*, 2022.

[165] Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. Signclip: Connecting text and sign language by contrastive learning. *arXiv preprint arXiv:2407.01264*, 2024.

[166] Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. Internal value alignment in large language models through controlled value vector activation. *arXiv preprint arXiv:2507.11316*, 2025.

[167] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

[168] Kristen Johnson and Dan Goldwasser. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 720–730, 2018.

[169] Kristen Johnson and Dan Goldwasser. Predicting moral values in political tweets. In *Proceedings of ...*, 2018.

[170] Larry Emerson Johnson and Sherif Rashad. An innovative system for real-time translation from american sign language (asl) to spoken english using a large language model (llm). In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 605–611. IEEE, 2024.

[171] Loso Judijanto, Alim Hardiansyah, and Opan Arifudin. Ethics and security in artificial intelligence and machine learning: Current perspectives in computing. *International Journal of Society Reviews (INJOSER)*, 3(2):374–380, 2025.

[172] Aishwarya Kamath and Rajarshi Das. A survey on semantic parsing. *arXiv preprint arXiv:1812.00978*, 2018.

[173] Zhehan Kang. Spoken language to sign language translation system based on hamnosys. In *Proceedings of the 2019 international symposium on signal processing systems*, pages 159–164, 2019.

[174] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. Moral concerns are differentially observable in language. *Cognition*, 212: 104696, 2021.

[175] Arif Ali Khan, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar. Ethics of ai: A systematic literature review of principles and challenges. In *Proceedings of the 26th international conference on evaluation and assessment in software engineering*, pages 383–392, 2022.

[176] Younas Khan, David Sánchez, and Josep Domingo-Ferrer. Federated learning-based natural language processing: a systematic literature review. *Artificial Intelligence Review*, 57(12):320, 2024.

[177] Peter Kieseberg, Edgar Weippl, A Min Tjoa, Federico Cabitza, Andrea Campagner, and Andreas Holzinger. Controllable ai-an alternative to trustworthiness in complex ai systems? In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 1–12. Springer, 2023.

[178] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 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, 2022.

[179] Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, 2023.

[180] Kimon Kieslich, Birte Keller, and Christopher Starke. Artificial intelligence ethics by design. evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, 9(1):20539517221092956, 2022.

[181] Hakpyeong Kim and Taehoon Hong. Emotion-oriented recommender system for personalized control of indoor environmental quality. *Building and Environment*, 254:111396, 2024.

[182] Hankyung Kim and Youn-kyung Lim. Teaching-learning interaction: a new concept for interaction design to support reflective user agency in intelligent systems. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1544–1553, 2021.

[183] Minbeom Kim, Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee, and Kyomin Jung. Critic-guided decoding for controlled text generation. *arXiv preprint arXiv:2212.10938*, 2022.

[184] Oleksandr Klymenko, Sebastian Meisenbacher, and Florian Matthes. Differential privacy in natural language processing: A survey. *arXiv*

preprint arXiv:2208.08140, 2022. URL `https://arxiv.org/abs/2208.08140`.

[185] Jonathan Kobbe, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*. Association for Computational Linguistics, ACL, 2020.

[186] Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 868–876. Association for Computational Linguistics, 2007.

[187] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[188] Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security*, 64:122–134, 2017.

[189] Tomoko Komatsu, Marisela Gutierrez Lopez, Stephann Makri, Colin Porlezza, Glenda Cooper, Andrew MacFarlane, and Sondess Missaoui. Ai should embody our values: Investigating journalistic values to inform ai technology design. In *Proceedings of the 11th Nordic conference on human-computer interaction: Shaping experiences, shaping society*, pages 1–13, 2020.

[190] X Kong, S Liu, and L Zhu. Toward human-centered xai in practice: a survey. mach. intell. res. 21, 740–770 (2024).

[191] Terry Koo, Frederick Liu, and Luheng He. Automata-based constraints for language model decoding. *arXiv preprint arXiv:2407.08103*, 2024.

[192] Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE, 2017.

[193] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

[194] Sebastian Laacke, Regina Mueller, Georg Schomerus, and Sabine Salloch. Artificial intelligence, social media and depression. a new concept of health-related digital autonomy. *The American Journal of Bioethics*, 21(7):4–20, 2021.

[195] Arto Laitinen and Otto Sahlgren. Ai systems and respect for human autonomy. *Frontiers in artificial intelligence*, 4:705164, 2021.

[196] Huije Lee, Jung-Ho Kim, Eui Jun Hwang, Jaewoo Kim, and Jong C Park. Leveraging large language models with vocabulary sharing for sign language translation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE, 2023.

[197] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for

algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–35, 2019.

[198] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model, 2024. URL `https://www.mixedbread.ai/blog/mxbai-embed-large-v1`.

[199] Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. Emona: Event-level moral opinions in news articles. *arXiv preprint arXiv:2404.01715*, 2024.

[200] Bruno Lepri, Nuria Oliver, and Alex Pentland. Ethical machines: The human-centric use of artificial intelligence. *IScience*, 24(3), 2021.

[201] John Kalung Leung, Igor Griva, William G Kennedy, Jason M Kinser, Sohyun Park, and Seo Young Lee. The application of affective measures in text-based emotion aware recommender systems. *arXiv preprint arXiv:2305.04796*, 2023.

[202] Jamy Li and Mark Chignell. Fmea-ai: Ai fairness impact assessment using failure mode and effects analysis. *AI and Ethics*, 2(4):837–850, 2022.

[203] Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

[204] Zelong Li, Wenyue Hua, Hao Wang, He Zhu, and Yongfeng Zhang. Formal-llm: Integrating formal language and natural language for controllable llm-based agents. *arXiv preprint arXiv:2402.00798*, 2024.

[205] Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al.

Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*, 2024.

[206] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

[207] Ji Hun Lim and Hun Yeong Kwon. A study on the modeling of major factors for the principles of ai ethics. In *Proceedings of the 22nd Annual International Conference on Digital Government Research*, pages 208–218, 2021.

[208] JongYoon Lim, Inkyu Sa, Bruce MacDonald, and Ho Seok Ahn. A sign language recognition system with pepper, lightweight-transformer, and llm. *arXiv preprint arXiv:2309.16898*, 2023.

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

[210] Elia Liscio and Simone Paolo Ponzetto. Mftc: Moral foundations twitter corpus for analyzing cross-domain morality in text. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, 2022.

[211] Elia Liscio, Dong Nguyen, Giulia Liscio, and Simone Paolo Ponzetto. Tomea: An explainable approach to supervised cross-domain moral rhetoric classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[212] Shaobo Liu, Guiran Liu, Binrong Zhu, Yuanshuai Luo, Linxiao Wu, and Rui Wang. Balancing innovation and privacy: Data security strategies in natural language processing applications. In *2024 5th*

International Conference on Machine Learning and Computer Application (ICMLCA), pages 609–613. IEEE, 2024.

[213] Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. Scope: Sign language contextual processing with embedding from llms. *arXiv preprint arXiv:2409.01073*, 2024.

[214] Qinghua Lu, Liming Zhu, Xiwei Xu, and Jon Whittle. Responsible-ai-by-design: a pattern collection for designing responsible ai systems. *arXiv preprint arXiv:2203.00905*, 2022.

[215] Jonas Lundberg, Mattias Arvola, and Karljohan Lundin Palmerius. Human autonomy in future drone traffic: Joint human–ai control in temporal cognitive work. *Frontiers in Artificial Intelligence*, 4:704082, 2021.

[216] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[217] Batnasan Luvaanjalba and Bo-chiuan Su. An epistemological analysis of the "brain in a vat" approach for the philosophy of artificial intelligence. In *International Conference on Human-Computer Interaction*, pages 97–111. Springer, 2022.

[218] Joseph B Lyons, Kevin T Wynne, Sean Mahoney, and Mark A Roebke. Trust and human-machine teaming: A qualitative study. In *Artificial intelligence for the internet of everything*, pages 101–116. Elsevier, 2019.

[219] Darshini Mahendran, Changqing Luo, and Bridget T Mcinnes.

Privacy-preservation in the context of natural language processing. *IEEE access*, 9:147600–147612, 2021.

[220] Musa Malik, Sungbin Youk, Frederic R Hopp, Oliver Scott Curry, Marc Cheong, Mark Alfano, and René Weber. The extended morality as cooperation dictionary (emacd): A crowd-sourced approach via the moral narrative analyzer platform. *Communication Methods and Measures*, pages 1–31, 2025.

[221] Theodora A Maniou and Andreas Veglis. Employing a chatbot for news dissemination during crisis: Design, implementation and evaluation. *Future Internet*, 12(7):109, 2020.

[222] Alessandro Mantelero and Maria Samantha Esposito. An evidence-based methodology for human rights impact assessment (hria) in the development of ai data-intensive systems. *Computer Law & Security Review*, 41:105561, 2021.

[223] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. Putting ai ethics into practice: The hourglass model of organizational ai governance. *arXiv preprint arXiv:2206.00335*, 2022.

[224] Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th international conference on computational linguistics*, pages 3659–3668, 2022.

[225] Lina Markauskaite, Rebecca Marrone, Oleksandra Poquet, Simon Knight, Roberto Martinez-Maldonado, Sarah Howard, Jo Tondeur, Maarten De Laat, Simon Buckingham Shum, Dragan Gašević, et al. Rethinking the entwinement between artificial intelligence and human

learning: What capabilities do learners need for a world with ai? *Computers and Education: Artificial Intelligence*, 3:100056, 2022.

[226] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[227] Tadahiro Matsumoto, Mihoko Kato, and Takashi Ikeda. Jspad: A sign language writing tool using signwriting. In *proceedings of the 3rd international universal communication symposium*, pages 363–367, 2009.

[228] Michael McCurrie and et al. Moral and affective film set. In *Proceedings of ...*, 2018.

[229] Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tür. Using in-context learning to improve dialogue safety. *arXiv preprint arXiv:2302.00871*, 2023.

[230] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[231] AI @Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

[232] Milagros Miceli, Julian Posada, and Tianling Yang. Studying up machine learning data: Why talk about bias when we mean power? *Pro-*

ceedings of the ACM on Human-Computer Interaction, 6(GROUP): 1–14, 2022.

[233] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[234] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507, 2019.

[235] Jakub Mlynar, Farzaneh Bahrami, André Ourednik, Nico Mutzner, Himanshu Verma, and Hamed Alavi. Ai beyond deus ex machina–reimagining intelligence in future cities with urban experts. In *Proceedings of the 2022 Chi conference on human factors in computing systems*, pages 1–13, 2022.

[236] Judith Molka-Danielsen, Jazz Rasool, and Carl H Smith. Design and deployment considerations for ethically advanced technologies for human flourishing in the workplace. In *IFIP working conference on human work interaction design*, pages 101–122. Springer, 2021.

[237] Amit Moryossef and Zifan Jiang. Signbank+: Multilingual sign language translation dataset. *arXiv preprint arXiv:2309.11566*, 2023.

[238] Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. Linguistically motivated sign language segmentation. *arXiv preprint arXiv:2310.13960*, 2023.

[239] Noor Nafees, Muhammad Azam, Aneesa Sohail, and Qaiser Janjua. Exploring the integration of ai for social-emotional learning: A psycho-

logical, technological, and educational approach. *The Critical Review of Social Sciences Studies*, 3(2):810–827, 2025.

[240] Thiloshon Nagarajah and Guhanathan Poravi. A review on automated machine learning (automl) systems. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE, 2019.

[241] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[242] An Nguyen, Li Chen, and Tanishq Saha. Cross-domain moral value classification with supervised neural models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[243] Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. Measuring moral dimensions in social media with mformer. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1134–1147, 2024.

[244] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric D Ragan, and Vibhav Gogate. On the importance of user backgrounds and impressions: Lessons learned from interactive ai applications. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–29, 2022.

[245] José Luiz Nunes, Guilherme FCF Almeida, Marcelo de Araujo, and

Simone DJ Barbosa. Are large language models moral hypocrites? a study based on moral foundations. *arXiv preprint arXiv:2405.11100*, 2024.

[246] Adrián Núñez-Marcos, Olatz Perez-de Viñaspre, and Gorka Labaka. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993, 2023.

[247] Humphrey O Obie, Waqar Hussain, Xin Xia, John Grundy, Li Li, Burak Turhan, Jon Whittle, and Mojtaba Shahin. A first look at human values-violation in app reviews. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 29–38. IEEE, 2021.

[248] Uran Oh, Hwayeon Joh, and YunJung Lee. Image accessibility for screen reader users: A systematic review and a road map. *Electronics*, 10(8):953, 2021.

[249] Abiodun Finbarrs Oketunji, Muhammad Anas, and Deepthi Saina. Large language model (llm) bias index–llmbi. *arXiv preprint arXiv:2312.14769*, 2023.

[250] Kirsten Ostherr. Artificial intelligence and medical humanities. *Journal of Medical Humanities*, 43(2):211–232, 2022.

[251] Achraf Othman and Mohamed Jemni. English-asl gloss parallel corpus 2012: Aslg-pc12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*, 2012.

[252] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina

Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[253] Sharon Oviatt. Technology as infrastructure for dehumanization: three hundred million people with the same face. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 278–287, 2021.

[254] Maria Pacheco and et al. Moral framing of covid-19 vaccination on twitter. In *Proceedings of ...*, 2022.

[255] Ilias Papastratis, Christos Chatzikonstantinou, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. Artificial intelligence technologies for sign language. *Sensors*, 21(17):5843, 2021.

[256] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. How accurate does it feel?–human perception of different types of classification mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[257] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[258] Princy Pappachan, Massoud Moslehpour, Ritika Bansal, and Mosiur Rahaman. Transparency and accountability. In *Challenges in large language model development and AI ethics*, pages 178–211. IGI Global Scientific Publishing, 2024.

[259] Pradeep Paraman and Sanmugam Anamalah. Ethical artificial intelligence framework for a good ai society: principles, opportunities and perils. *AI & SOCIETY*, 38(2):595–611, 2023.

[260] Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. *arXiv preprint arXiv:2405.21047*, 2024.

[261] Kanghee Park, Timothy Zhou, and Loris D'antoni. Flexible and efficient grammar-constrained decoding. 2025. URL `https://api.semanticscholar.org/CorpusID:276235743`.

[262] Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, 57(6):1–40, 2025.

[263] Silviu Paun, Ron Artstein, and Massimo Poesio. *Statistical methods for annotation analysis*. Springer Nature, 2022.

[264] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

[265] Vjosa Preniqi, Iacopo Ghinassi, Kyriaki Kalimeri, and Charalampos Saitis. Moralbert: Detecting moral values in social discourse. *arXiv preprint arXiv:2403.07678*, 2024.

[266] Siegmund Prillwitz and Heiko Zienert. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Current trends in European Sign Language Research. Proceed-*

ings of the 3rd European Congress on Sign Language Research, pages 355–379, 1990.

[267] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.

[268] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[269] Aung Pyae. What is human-centeredness in human-centered ai? development of human-centeredness framework and ai practitioners' perspectives. *arXiv preprint arXiv:2502.03293*, 2025.

[270] Valentina Pyatkin, Frances Yung, Merel CJ Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. Design choices for crowdsourcing implicit discourse relations: revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032, 2023.

[271] Honglin Qin, Hongye Zheng, Bingxing Wang, Zhizhong Wu, Bingyao Liu, and Yuanfang Yang. Reducing bias in deep learning optimization: The rsgdm approach. In *2024 4th International Conference on Computer Science and Blockchain (CCSB)*, pages 99–103. IEEE, 2024.

[272] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. Valuenet: A new dataset for human value

driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11183–11191, 2022.

[273] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[274] Alessio Ragno, Biagio La Rosa, and Roberto Capobianco. Prototype-based interpretable graph neural networks. *IEEE Transactions on Artificial Intelligence*, 5(4):1486–1495, 2022.

[275] Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.

[276] Inioluwa Deborah Raji, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jutta Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44. ACM, 2020. doi: 10.1145/3351095.3372873. URL `https://dl.acm.org/doi/10.1145/3351095.3372873`.

[277] Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*, 2024.

[278] S. Ray and et al. Benchmarking human-centered ai systems. In *Pro-*

ceedings of the 2023 AAAI Conference on Human Computation and Crowdsourcing, 2023.

[279] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `http://arxiv.org/abs/1908.10084`.

[280] Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. *arXiv preprint arXiv:2406.04214*, 2024.

[281] André Renz and Gergana Vladova. Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, 11(5), 2021.

[282] Jimin Rhim, Ji-Hyun Lee, Mo Chen, and Angelica Lim. A deeper look at autonomous vehicle ethics: an integrative ethical decision-making framework to explain moral pluralism. *Frontiers in Robotics and AI*, 8:632394, 2021.

[283] Dalai Dos Santos Ribeiro, Gabriel Diniz Junqueira Barbosa, Marisa Do Carmo Silva, Hélio Lopes, and Simone Diniz Junqueira Barbosa. Exploring the impact of classification probabilities on users' trust in ambiguous instances. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–9. IEEE, 2021.

[284] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceed-*

ings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144, New York, NY, USA, 2016. ACM.

[285] Mark O Riedl. Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies*, 1(1):33–36, 2019.

[286] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[287] Christina Rödel, Susanne Stadler, Alexander Meschtscherjakov, and Manfred Tscheligi. Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. In *Proceedings of the 6th international conference on automotive user interfaces and interactive vehicular applications*, pages 1–8, 2014.

[288] Kat Roemmich and Nazanin Andalibi. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.

[289] Andrew Rojecki and et al. Moral foundations and the covid-19 pandemic: Annotated twitter corpus. 2021.

[290] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user

studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122, 2023.

[291] Shamik Roy and Dan Goldwasser. Extending moral foundations annotation to us congress tweets. In *Proceedings of ...*, 2021.

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

[293] Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[294] Mark Ryan and Bernd Carsten Stahl. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, 19(1):61–86, 2021.

[295] P Sam Sahil, Anupam Jamatia, and Kunal Chakma. Synergizing contextual semantics and moral knowledge graphs: A dual-path architecture for moral foundation prediction.

[296] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[297] Supraja Sankaran and Panos Markopoulos. " it's like a puppet master": User perceptions of personal autonomy when interacting with intelligent technologies. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, pages 108–118, 2021.

[298] Supraja Sankaran, Chao Zhang, Henk Aarts, and Panos Markopoulos. Exploring peoples' perception of autonomy and reactance in everyday ai interactions. *Frontiers in psychology*, 12:713074, 2021.

[299] Manish Sanwal. Layered chain-of-thought prompting for multi-agent llm systems: A comprehensive approach to explainable large language models. *arXiv preprint arXiv:2501.18645*, 2025.

[300] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

[301] Laura Sartori and Andreas Theodorou. A sociotechnical perspective for the future of ai: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1):4, 2022.

[302] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: expertise and explanations. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 240–251, 2019.

[303] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

[304] Daniel Schiff, Aladdin Ayesh, Laura Musikanski, and John C Havens. Ieee 7010: A new standard for assessing the well-being implications of artificial intelligence. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 2746–2753. IEEE, 2020.

[305] Stefan Schmager, Ilias Pappas, and Polyxeni Vassilakopoulou. Defining human-centered ai: a comprehensive review of hcai literature. 2023.

[306] Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. *arXiv preprint arXiv:2109.05093*, 2021.

[307] Daniel Schroter, Daryna Dementieva, and Georg Groh. Adam-smith at semeval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. *arXiv preprint arXiv:2305.08625*, 2023.

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

[309] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P Dickerson. Group fairness in bandits with biased feedback. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, 2022.

[310] Shalom H Schwartz. A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey*, 259(290):261, 2003.

[311] Shalom H Schwartz. Basic human values: Theory, methods, and application. *Risorsa Uomo*, (2007/2), 2007.

[312] Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. Extending the cross-cultural validity

of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, 32(5):519–542, 2001.

[313] Shalom H Schwartz et al. Basic human values: An overview. 2006.

[314] Luciano Serafini, Raul Barbosa, Jasmin Grosinger, Luca Iocchi, Christian Napoli, Salvatore Rinzivillo, Jacques Robin, Alessandro Saffiotti, Teresa Scantamburlo, Peter Schüller, et al. On some foundational aspects of human-centered artificial intelligence. *arXiv preprint arXiv:2112.14480*, 2021.

[315] Antonio FG Sevilla, Alberto Díaz Esteban, and José María Lahoz-Bengoechea. Automatic signwriting recognition: Combining machine learning and expert knowledge to solve a novel problem. *IEEE Access*, 11:13211–13222, 2023.

[316] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2020.102551. URL `https://www.sciencedirect.com/science/article/pii/S1071581920301531`.

[317] Donghee Shin, Kerk F. Kee, and Emily Y. Shin. Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management*, 65:102494, 2022. ISSN 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2022.102494. URL `https://www.sciencedirect.com/science/article/pii/S0268401222000251`.

[318] Richard Shin, Christopher H Lin, Sam Thomson, Charles Chen,

Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. Constrained language models yield few-shot semantic parsers. *arXiv preprint arXiv:2104.08768*, 2021.

[319] Ben Shneiderman. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Trans. Interact. Intell. Syst.*, 10(4), October 2020. ISSN 2160-6455. doi: 10.1145/3419764. URL `https://doi.org/10.1145/3419764`.

[320] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020. doi: 10.1080/10447318.2020.1741118. URL `https://doi.org/10.1080/10447318.2020.1741118`.

[321] Ben Shneiderman. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, pages 109–124, 01 2020. doi: 10.17705/1thci.00131.

[322] Ben Shneiderman. Human-centered ai: A new synthesis. In *IFIP Conference on Human-Computer Interaction*, pages 3–8. Springer, 2021.

[323] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[324] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011.

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

[326] Nathalie A Smuha. The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4): 97–106, 2019.

[327] Harri Söderholm, Nina Nikki, Zdeňka Špiclová, and Jimi Vesala. Seek first the kingdom of cooperation: Testing the applicability of morality-as-cooperation theory to the sermon on the mount. *Open Theology*, 11(1):20250040, 2025.

[328] HAI Stanford. Stanford institute for human-centered artificial intelligence, 2020.

[329] Christina E Stimson and Rebecca Raper. Participatory ai: a method for integrating inclusive and ethical design considerations into autonomous system development. In *Annual Conference Towards Autonomous Robotic Systems*, pages 144–154. Springer, 2024.

[330] Marco Stranisci and et al. Moralconvita: An italian twitter corpus for moral foundations analysis. In *Proceedings of ...*, 2021.

[331] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[332] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[333] Valerie Sutton. *Lessons in SignWriting*. SignWriting Press, 2022.

[334] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). 2023.

[335] Mokanarangan Thayaparan, Marco Valentino, and André Freitas. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*, 2020.

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

[337] Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*, 2022.

[338] John Trager and et al. Moral foundations reddit corpus. In *Proceedings of ...*, 2022.

[339] Gabriele Tuccio, Luana Bulla, Maria Madonia, Aldo Gangemi, and Misael Mongiovì. GRAMMAR-LLM: Grammar-constrained natural language generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3412–3422, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.177. URL `https://aclanthology.org/2025.findings-acl.177/`.

[340] Steven Umbrello and Ibo Van de Poel. Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, 1(3):283–296, 2021.

[341] Michiel van der Meer, William Held, Jonathan Gordon, and Malvina Nissim. Annotator-centric active learning: Capturing diverse human perspectives in subjective nlp tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[342] Bart van Luenen. Cross-domain detection of moral rhetoric in extremist texts. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020.

[343] Juan Vargas and et al. Mftcxplain: A multilingual moral foundations hate speech corpus with explanations. 2025.

[344] Susana M Vieira, Uzay Kaymak, and João MC Sousa. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems*, pages 1–8. IEEE, 2010.

[345] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[346] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7 (2):76–99, 2017. doi: 10.1093/idpl/ipx005. URL `https://academic.oup.com/idpl/article/7/2/76/3860948`.

[347] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander

Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, November 2019. ISSN 2573-0142. doi: 10.1145/3359313. URL http://dx.doi.org/10.1145/3359313.

[348] Dakuo Wang, Josh Andres, Justin D. Weisz, Erick Oduor, and Casey Dugan. Autods: Towards human-centered automation of data science. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pages Article 79, 1–12, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3411764.3445526.

[349] Hao Wang and Vincent Blok. Why putting artificial intelligence ethics into practice is not enough: Towards a multi-level framework. *Big Data & Society*, 12(2):20539517251340620, 2025.

[350] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445645. URL https://doi.org/10.1145/3411764.3445645.

[351] Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *arXiv preprint arXiv:2406.14194*, 2024.

[352] Nathaniel Weber and et al. Moral foundations news corpus. In *Proceedings of ...*, 2021.

[353] Brandon T Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

[354] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[355] Alan FT Winfield and Marina Jirotka. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180085, 2018.

[356] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation. *arXiv preprint arXiv:2405.04164*, 2024.

[357] Jing Yi Xie, Graeme Hirst, and Yang Xu. Contextualized moral inference. *arXiv preprint arXiv:2008.10762*, 2020.

[358] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pages Article 83, 1–16, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3411764.3445306.

[359] Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, 26(4):42–46, 2019.

[360] Wei Xu and Zaifeng Gao. Enabling human-centered ai: A methodological perspective. *arXiv preprint arXiv:2311.06703*, 2023.

[361] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human–Computer Interaction*, 39(3):494–518, 2022. doi: 10.1080/10447318.2022.2041900. URL `https://doi.org/10.1080/10447318.2022.2041900`.

[362] Feng Yang, Ghada Zamzmi, Sandeep Angara, Sivaramakrishnan Rajaraman, André Aquilina, Zhiyun Xue, Stefan Jaeger, Emmanouil Papagiannakis, and Sameer K Antani. Assessing inter-annotator agreement for medical image segmentation. *IEEE Access*, 11:21300–21312, 2023.

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

[364] Jing Yao, Xiaoyuan Yi, and Xing Xie. Clave: An adaptive framework for evaluating values of llm generated responses. *Advances in Neural Information Processing Systems*, 37:58868–58900, 2024.

[365] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[366] Kayo Yin and Jesse Read. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*, 2020.

[367] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025.

[368] Lorenzo Zangari, Candida M Greco, Davide Picca, and Andrea Tagarelli. Me2-bert: Are events and emotions what you need for moral foundation prediction? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9516–9532, 2025.

[369] Lorenzo Zangari, Candida Maria Greco, Davide Picca, and Andrea Tagarelli. A survey on moral foundation theory and pre-trained language models: Current advances and challenges. *AI & SOCIETY*, pages 1–26, 2025.

[370] Honghua Zhang, Meihua Dang, Nanyun Peng, and Guy Van den Broeck. Tractable control for autoregressive language generation. In *International Conference on Machine Learning*, pages 40932–40945. PMLR, 2023.

[371] Honghua Zhang, Po-Nien Kung, Masahiro Yoshida, Guy Van den Broeck, and Nanyun Peng. Adaptable logical control for large language models. *arXiv preprint arXiv:2406.13892*, 2024.

[372] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[373] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

[374] Hongye Zheng, Bingxing Wang, Minheng Xiao, Honglin Qin, Zhizhong Wu, and Lianghao Tan. Adaptive friction in deep learning: Enhancing optimizers with sigmoid and tanh function. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 809–813. IEEE, 2024.

[375] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881, 2023.

[376] Yuqiu Zhou, Wei He, Weizhen Hou, and Ying Zhu. Pianno: a probabilistic framework automating semantic annotation for spatial transcriptomics. *Nature Communications*, 15(1):2848, 2024.

[377] Wenhao Zhu, Yuhang Xie, Guojie Song, and Xin Zhang. Eavit: Efficient and accurate human value identification from text data via llms. *arXiv preprint arXiv:2505.12792*, 2025.

[378] Caleb Ziems and et al. Moral integrity corpus: Extending social chemistry 101 with moral annotations. 2022.

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

[380] Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. Llms for xai: Future directions for explaining explanations. *arXiv preprint arXiv:2405.06064*, 2024.