Warning: The appendix includes explanations of value concepts, which are dual-sided. It contains negative examples that can be toxic, upsetting, or offensive.

A Introduction to the Explored Values

Given that the concepts we delve into are inherently rooted in ethics and morals, it's essential to clarify their ethical foundations. Below, we present the ethical theory as summarized by Vida et al. (2023). Grounded in this theoretical framework, we then elucidate the definitions and ethical characteristics of each value we explore.

A.1 Ethical Theory

According to Vida et al. (2023), Ethics is divided into four branches: *normative ethics*, *applied ethics*, *descriptive ethics*, and *metaethics*.

Specifically, *normative ethics* focuses on the principles and criteria that define moral correctness. It operates within a framework of universal norms and values, providing justification for what is deemed right or wrong. *Descriptive ethics*, conversely, involves empirical investigations to describe or explain the moral judgments, preferences, and value systems prevalent in societies. It refrains from making moral judgments, focusing instead on documenting and analyzing prevailing ethical beliefs and behaviors. *Applied ethics* extends the general norms and values from *normative ethics* to specific contexts and fields, dealing with concrete ethical dilemmas and decisions in domains like bioethics, environmental ethics, or, as relevant to our paper, the ethics of artificial intelligence. *Metaethics* lays the analytical foundation for these three branches, delving into the nature of moral language, the meaning of moral judgments, and the foundational aspects of ethical theories.

Furthermore, *normative ethics* can be assigned to three competing ethical families: *virtue ethics*, *deontological ethics*, and *consequentialism*. While *deontological ethics* emphasizes the intrinsic rightness or wrongness of actions based on principles or rules, *consequentialism* assesses actions by their outcomes or consequences. Meanwhile, *virtue ethics* focuses on the moral character and virtues of the individual.

A.2 Definitions and Ethical Characteristics of Each Value

Below, we detail the definitions of the 7 explored values, their ethical characteristics, and any interconnections between them.

Commonsense Morality Commonsense Morality refers to the intuitive and widely accepted moral principles guiding everyday human behavior. These principles often stem from societal norms, cultural values, and emotional responses, forming the basis of our ethical decision-making. Commonsense Morality focuses on evaluating actions based on moral correctness rather than merely describing existing moral beliefs and behaviors in society. Thus, it can be categorized as a part of *normative ethics*.

Deontology Deontology, on the other hand, focuses on the inherent rightness or wrongness of actions based on adherence to a set of rules or constraints. It asserts that certain actions possess moral obligations or prohibitions, independent of their outcomes. Thus, Deontology is categorized under *normative ethics*, specifically within the *deontological ethics* family. While both Commonsense Morality and Deontology belong to *normative ethics*, they differ in their foundational principles. Commonsense Morality is anchored in societal norms and moral correctness, emphasizing the alignment of actions with shared societal values. In contrast, Deontology prioritizes rule-based morality, focusing on the inherent moral obligations or prohibitions associated with actions, regardless of their outcomes.

Utilitarianism Utilitarianism emphasizes maximizing overall well-being, aiming for a world where every individual experiences the highest possible level of well-being. Belonging to the *consequentialism* family within *normative ethics*, utilitarianism assesses the moral value of an action based on its outcomes or consequences, contrasting with deontology's focus on the intrinsic rightness or wrongness of actions.

Fairness Fairness pertains to the equitable and impartial treatment of individuals, regardless of their demographic attributes such as race, gender, age, religion, or socioeconomic status. Its emphasis on societal biases places Fairness within the realm of *descriptive ethics*, focusing less on absolute moral rightness or wrongness.

Truthfulness Truthfulness involves the accurate representation of facts about the real world. In this context, a statement is considered truthful if it aligns with objective reality, without being influenced by personal beliefs or biases. Given that