What Should (A)I Do? The Heuristic Role of Artificial Advice in Moral Decision-Making

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Fakultät für Humanwissenschaften der Universität Regensburg

vorgelegt von

Julius Wolff

geboren in Gräfelfing

Regensburg 2026

Gutachter (Betreuer): Prof. Dr. Peter Fischer

Gutachter: Prof. Dr. Andreas Kastenmüller

Abstract

This dissertation investigates how Artificial Intelligence (AI) affects moral decision-making across four empirical studies, examining the extent, conditions, and mechanisms of AI- induced moral persuasion. Building on previous work, the overarching aim was to determine whether AI-generated advice affects moral decisions to the same extent as human advice, and whether advice generally functions as a source-independent heuristic cue in moral decision-making. Across four empirical studies, we systematically varied dilemma difficulty, ecological validity, social proximity, conflicting advice, and the temporal order of decision and advice to test the robustness and generalizability of artificial moral persuasion.

Study 1 provided the first systematic demonstration that AI-generated advice affects moral decision-making to a comparable extent as human advice. Across nine sacrificial moral dilemmas of varying difficulty, advice was presented with explanatory context rather than simple imperatives. Deontological advice reliably increased deontological choices, whereas utilitarian advice showed little persuasive effect. Importantly, AI- and human advice exerted comparable influence, indicating that advisor source did not meaningfully affect advice use even in complex, morally charged situations.

Study 2 replicated the findings of Study 1 and extended them to everyday moral decision-making while experimentally manipulating social proximity across twenty dilemmas. Both altruistic and egoistic advice significantly affected moral decisions, and AI-generated advice again proved to be as persuasive as human advice. Although social proximity influenced the overall strength of persuasion, it did not introduce systematic differences between advisors. These results suggest that everyday moral decision-making is similarly susceptible to advice, consistent with its use as a heuristic cue rather than as a source-dependent signal.

Study 3 further clarified the underlying psychological processes involved in the heuristic use of advice by presenting participants with simultaneous AI- and human advice that was either convergent or conflicting across sacrificial and everyday moral dilemmas. When advice was convergent, participants predominantly followed the recommended option; when advice conflicted, decisions were guided by advice content and perceived helpfulness rather than by advisor. This pattern indicates that both AI- and human advice are primarily used heuristically.

Study 4 extended the findings of the previous studies to situations involving pre-existing personal positions. Participants first made initial moral decisions and subsequently received AI- or human advice before deciding again. Decision stability declined substantially when advice opposed the initial choice, particularly for utilitarian judgments in sacrificial dilemmas and for everyday decisions more broadly. AI- and human advice again produced comparable effects, and perceived helpfulness reliably predicted whether participants revised their earlier decisions.

Across all four studies, no meaningful differences emerged between the persuasive impact of AI and human advisors. Instead, moral persuasion was consistently driven by normative alignment, subjective usefulness, and contextual features of the dilemma. These findings suggest that artificial advisors can function as potent sources of social influence and that individuals readily adopt AI-generated advice as an effort-reducing heuristic in moral decision-making. The dissertation concludes by discussing implications for moral psychology, human-AI interaction, and the societal challenges posed by increasingly advisory-capable AI systems.

Keywords: moral decision-making, Artificial Intelligence, advice, ethics

Zusammenfassung

Diese Dissertation untersucht im Rahmen vier empirischer Studien, wie KI-generierte Ratschläge das moralische Entscheidungsverhalten beeinflussen. Der Fokus liegt dabei darauf, in welchem Ausmaß, unter welchen Bedingungen und über welche Mechanismen KI-generierte moralische Ratschläge menschliches Entscheidungsverhalten beeinflussen. Aufbauend auf der bisherigen Forschung ist das übergeordnete Ziel, zu replizieren, ob KI-generierte Ratschläge einen signifikanten Einfluss auf moralische Entscheidungen haben und ob dieser mit dem Einfluss menschlicher Ratgeber vergleichbar ist. In den Studien werden die Schwierigkeit der Dilemmata, die ökologische Validität, die soziale Nähe und die zeitliche Abfolge von Ratschlag und Entscheidung systematisch variiert, um die Robustheit und Verallgemeinerbarkeit der Effekte der bisherigen Forschung zu testen.

Studie 1 repliziert und erweitert die Ergebnisse bisheriger Forschung, indem sie den Einfluss von KI-generierten Ratschlägen in neun moralischen Dilemmata unterschiedlicher Schwierigkeitsgrade untersucht. Die Ergebnisse deuten darauf hin, dass deontologische Ratschläge die Anzahl deontologischer Entscheidungen zuverlässig beeinflussen. Utilitaristische Ratschläge haben dagegen nur eine geringe Überzeugungskraft auf das Entscheidungsverhalten. Zwischen beiden Ratschlagsquellen lässt sich kein signifikanter Unterschied feststellen.

In Studie 2 werden die bisherigen Ergebnisse um die Dimension alltäglicher moralischer Konflikte erweitert und die soziale Distanz in zwanzig Dilemmata experimentell manipuliert. Sowohl altruistische als auch egoistische Ratschläge beeinflussen die Entscheidungen signifikant. Es lassen sich erneut keine Unterschiede zwischen KI-generierten und menschlichen Ratschläge feststellen. Die Ergebnisse zeigen, dass die soziale Nähe die Gesamtstärke der Überzeugungskraft beeinflusst, jedoch lassen sich auch bei der Manipulation der sozialen Nähe keine systematischen Unterschiede zwischen beiden Ratgebern erkennen.

Studie 3 baut auf den Ergebnissen von Studien 1 und 2 auf und untersucht den direkten Vergleich von KI-generierten und menschlichen Ratschlägen auf Moralentscheidungen. Im Rahmen der Studie werden gleichzeitig KI- und menschliche Empfehlungen präsentiert, die entweder übereinstimmend oder widersprüchlich sind. Bei übereinstimmenden Ratschlägen, folgen die Teilnehmer stark der empfohlenen Option. Bei widersprüchlichen Ratschlägen wird die Entscheidungen eher durch den Inhalt der Ratschläge und die wahrgenommene

Nützlichkeit als durch die Quelle des Ratschlags bestimmt. Eine wahrgenommene Ähnlichkeit mit dem menschlichen Ratgeber erhöht die Annahme menschlicher Ratschläge signifikant. Eine hohe Akzeptanz und eine wahrgenommene Nützlichkeit der KI-generierten Ratschläge sagt die Nutzung des KI Ratschlags signifiant voraus. Über alle Dilemmata und Bedingungen zeigt sich keine inhärente Präferenz gegenüber einer der beiden Quellen.

Studie 4 untersucht, ob sich der in den bisherigen Studien beobachtete Einfluss auch auf bereits getroffene Entscheidungen generalisieren lässt. Die Teilnehmer trafen zunächst moralische Entscheidungen und erhielten anschließend Ratschläge von einer KI oder einem Menschen, bevor sie erneut entscheiden mussten. Die Entscheidungsstabilität nimmt signifikant ab, wenn die Ratschläge im Widerspruch zur ursprünglichen Entscheidung stehen, insbesondere bei utilitaristischen Entscheidungen, in Dilemmata mit starken Konsequenzen und bei alltäglichen Entscheidungen im Allgemeinen. KI und menschliche Berater erzielten erneut vergleichbare Effekte, und die wahrgenommene Nützlichkeit sagte voraus, ob die Teilnehmer ihre früheren Entscheidungen revidierten.

Über alle vier Studien zeigten sich keine signifikanten Unterschiede zwischen der Überzeugungskraft von KI-generierten und menschlichen Ratschlägen. Stattdessen wird die moralische Überzeugungskraft durch normative Übereinstimmung, subjektive Nützlichkeit und kontextuelle Merkmale des Dilemmas beeinflusst. Die Ergebnisse deuten darauf hin, dass KI-basierte Ratgeber als starke soziale Einflussfaktoren fungieren können und, dass Menschen KI-generierte Ratschläge als Heuristik bei moralischen Urteilen annehmen, um die mentale Auseinandersetzung mit den moralischen Entscheidungen zu reduzieren. Die Dissertation schließt mit einer Diskussion der Implikationen für die Moralpsychologie, die Interaktion zwischen Mensch und KI und die gesellschaftlichen Herausforderungen, die sich aus zunehmend beratungsfähigen KI-Systemen ergeben.

Schlagwörter: Moralentscheidungen, Künstliche Intelligenz, Ratschläge, Ethik

Contents

List of Figures

List of Tables

1 General Introduction

Over the past decade, Artificial Intelligence (AI) has shifted from a predominantly analytical technology to an increasingly interactive and advisory presence in everyday life. While early computational systems were designed to optimize clearly defined parameters such as speed, efficiency, or accuracy, contemporary systems are capable of producing language, explanations, recommendations, and judgments that resemble human advice (Weng et al., 2024). As a result, Artificial Intelligence is no longer confined to the background of decision-making processes but increasingly appears in contexts that are social, interpretative, and morally laden.

This development marks a profound transformation in the cognitive ecology of human decision-making. Traditionally, moral decisions and normative guidance were embedded in interpersonal relationships, cultural practices, institutions, and internal reflection. Today, artificial systems are capable of entering this space by offering suggestions not only about how to act, but implicitly about what may be right, reasonable, or justifiable in a given situation. In doing so, they interact with some of the most fundamental aspects of human cognition: moral reasoning, responsibility attribution, trust, and agency (Crawford, 2021).

Research in moral psychology has long demonstrated that moral decision-making is not primarily the outcome of slow, deliberate reasoning. Instead, moral evaluations often arise from fast, intuitive, and emotionally grounded processes, followed by post-hoc rationalization (J. D. Greene et al., 2008; Haidt, 2001). Faced with complexity, ambiguity, or personal conflict, individuals tend to rely on cognitive shortcuts that reduce uncertainty and effort. Heuristics, in this sense, are not simply sources of bias, but adaptive mechanisms that enable efficient functioning in a world characterized by limited time, attention, and cognitive resources (Kahneman, 2011; Shah & Oppenheimer, 2008).

Advice represents one such shortcut. Social and informational input from others has long been known to shape perception, judgment, and behavior (Cialdini & Goldstein, 2004). Advice narrows the decision space, legitimizes specific choices, and can reduce both cognitive conflict

and emotional burden. In moral contexts, it can function as a normative anchor, a justification mechanism, or a corrective signal (Gigerenzer, 2008; Rand et al., 2014; Sunstein, 2005). In situations involving conflicts between competing moral and practical considerations, guidance may influence not only the decision itself but also how morally legitimate that decision is perceived to be. Increasingly, such guidance does not come exclusively from other people or institutional rules, but from artificial systems that process contextual information and generate suggestions based on inferred priorities and goals.

As artificial systems increasingly generate advice, they begin to occupy a novel position within this cognitive and social landscape. One can imagine this for example in everyday situations involving conflicting obligations, when an artificial assistant proposes how to prioritize between competing professional, social, or personal demands, and thereby implicitly frames what appears to be the most reasonable or morally appropriate course of action. Emerging research suggests that individuals often attribute algorithmic output with objectivity, consistency, and impartiality, a tendency referred to as the machine heuristic (Araujo et al., 2020). At the same time, research on algorithm aversion and algorithm appreciation highlights the conditional and context-dependent nature of trust in artificial systems (Castelo et al., 2019; Dietvorst et al., 2015; Logg et al., 2019). While some reject algorithmic input in domains perceived as subjective or value-laden, others actively prefer it, particularly in situations that are complex, uncertain, or associated with perceived human bias. Preliminary work indicates that these tendencies extend into the moral domain. AI moral advice can influence human decisions in ways comparable to human recommendations, even when users are fully aware of the non-human nature of the source (Krügel et al., 2022, 2023b, 2025). In some cases, exposure to self-interested or unethical artificial advice has been shown to increase corresponding behavior, revealing the dual potential of such systems to function either as moral scaffolds or as instruments of moral disengagement (Köbis et al., 2021; Leib et al., 2023). These findings suggest that artificial advisors do not simply assist decision-making; they can subtly reshape moral standards, responsibility attribution, and patterns of justification. Conceptually, this positions Artificial Intelligence not only as a tool, but as a new kind of social and cognitive actor. Recent work on Artificial Moral Advisors (AMAs) emphasizes that such systems need not be granted moral agency to exert moral influence. Instead, they operate as

external inputs that interact with human intuition, reasoning, and social cognition (Giubilini & Savulescu, 2018; Rodríguez-López & Rueda, 2023). Their impact therefore depends less on their internal architecture and more on how they are perceived, interpreted, and integrated into existing cognitive processes.

The present dissertation is situated at this intersection of moral psychology, heuristics research, and human–AI interaction. It investigates how AI advice is processed in morally relevant situations and under which psychological and situational conditions such advice shapes, reinforces, or destabilizes human moral judgment. The dissertation pursues two central objectives. First, it develops a theoretical account of artificial advice as a cognitive and moral heuristic, grounded in established models of human decision-making. Second, it presents a series of empirical studies designed to systematically examine how individuals react to AI-generated moral advice across varying contexts of complexity, emotional involvement, source conflict, and temporal distance. The following chapters take an interdisciplinary approach. Chapter 2.1 reviews the psychological foundations of morality, focusing on intuitive and deliberative processes and the role of social and contextual cues in moral decision-making. Chapter 2.2 examines Artificial Intelligence from a conceptual and technological perspective, tracing its development from computational tool to social and advisory system. Chapter 2.3 integrates these frameworks to delineate key research questions and introduces four empirical studies that investigate when, why, and how AI-generated moral advice influences human decision-making. In doing so, this dissertation seeks not only to describe an emerging phenomenon, but to anticipate its longer-term implications. As artificial systems continue to permeate domains traditionally governed by human ethical reasoning, understanding their psychological impact becomes essential. The central question is no longer whether Artificial Intelligence will participate in moral decision-making, but how human moral agency can be preserved, strengthened, and guided in a world increasingly shaped by algorithmic counsel.

1.1 Research Problem and Objectives

As AI systems increasingly participate in ethically relevant decision-making, a new psychological and societal question emerges: how do humans integrate, resist, or reinterpret moral advice when it originates from a non-human, artificial source? Although previous work

demonstrates that artificial advice can meaningfully influence moral decisions (Krügel et al., 2022, 2023b, 2025), the conditions, mechanisms, and limits of this influence remain only partially understood. In particular, it is still unclear how different moral contexts, relational dynamics, and temporal dynamics shape the reception of artificial moral advice.

Drawing on theories from moral psychology, social influence, and human–AI interaction, the research seeks to identify the conditions under which AI advisors are accepted as legitimate sources of guidance, and when such reliance may compromise independent moral reflection and responsibility. In doing so, the dissertation contributes to an emerging psychological understanding of how trust, authority, and accountability are negotiated between human actors and intelligent systems in morally significant situations.

Four interrelated research objectives guide the empirical program:

- **Robustness of AI influence.** The first objective is to test whether the persuasive impact of AI-generated moral advice generalizes across different types of dilemmas and levels of complexity, thereby establishing the stability of previously observed effects under varying moral conditions.
- **Social proximity and emotional context.** The second objective is to examine how relational closeness and emotional relevance shape receptivity to AI advice by contrasting altruistic and egoistic everyday dilemmas across varying degrees of social distance, extending research beyond abstract sacrificial dilemmas.
- **Source conflict and comparative weighting.** The third objective is to investigate how individuals respond when human and artificial advisors provide convergent or conflicting advice, revealing potential preferences, biases, or parity in how different advisors are weighted.
- **Temporal stability and retrospective change.** The fourth objective is to explore how AI advice interacts with initial, intuitive moral decisions over time, assessing whether such advice consolidates, shifts, or destabilizes prior decisions during post-decisional reflection.

Together, these objectives define a cohesive, theory-driven research agenda examining whether and under what conditions AI shapes moral decision-making.

1.2 Overview of the Dissertation

This dissertation comprises four empirical studies that together examine the psychological mechanisms, contextual moderators, and boundary conditions of AI-generated moral advice. Each study addresses a distinct, yet interrelated, aspect of human–AI interaction, enabling an integrated understanding of how artificial advisors shape moral reasoning across different contexts of complexity, social embeddedness, and temporal perspective.

Study 1 examines the robustness of AI moral influence. By systematically varying the structure, difficulty, and complexity of multiple moral dilemmas, the study tests whether the persuasive impact of AI advice generalizes across moral contexts. It extends previous findings (Krügel et al., 2023b, 2025) by replicating the effect across a broader range of scenarios and by examining whether advice transparency enhances reflective endorsement.

Study 2 examines the role of social proximity in advice reception by introducing everyday moral dilemmas that contrast altruistic and egoistic responses and manipulate the relational distance between the decision-maker and affected individuals. This design enhances ecological validity and tests whether empathic engagement weakens or amplifies reliance on artificial advisors.

Study 3 focuses on the integration of advice under conditions involving multiple sources. It introduces a dual-source paradigm in which participants receive either convergent or conflicting advice from a human and an AI advisor. By comparing advice weighting across source types, the study examines whether individuals display advisor preferences, systematic biases, or parity in how human and AI advice are incorporated.

Study 4 adds a temporal dimension by examining post-decisional change. The study analyzes how AI advice influences decisions after an initial choice has been made. It assesses whether AI advice stabilizes, alters, or destabilizes initial moral decisions, thereby providing insight into the durability and dynamics of artificial moral influence over time.

Together, these studies form a cumulative and coherent research program that combines experimental control with ecological realism. They elucidate when and why individuals follow or resist AI-generated moral advice, how emotional and contextual factors modulate its impact, and to what extent artificial systems can assume enduring advisory roles in morally significant decisions.

1.3 Structure of the Dissertation

This monograph is organized to provide both conceptual depth and empirical integration. Following the present general introduction, chapter 2.1 reviews the theoretical foundations of moral psychology, tracing the development of moral reasoning from philosophical approaches to contemporary models such as the Social Intuitionist Model and Moral Foundations Theory. This chapter establishes how moral decisions emerge through the interaction of intuitive and deliberative processes and how external input, including advice, can shape this interplay.

Chapter 2.2 then turns to Artificial Intelligence, outlining its conceptual scope, historical development, and increasing role as a social and advisory system. This chapter situates AI within the broader context of decision-making and social cognition, emphasizing its transformation from analytical tool to autonomous advisory agent capable of influencing human decision-making in ethically significant ways.

Chapter 2.3 integrates insights from moral psychology and AI research to articulate the overarching conceptual framework guiding the dissertation. It delineates key theoretical tensions and empirical challenges associated with AI-generated moral advice and situates the present research within this interdisciplinary landscape.

Chapters 3 through 6 present the four empirical studies in detail. Each chapter includes the relevant theoretical background, methodological approach, statistical analyses, and interpretation of results, with explicit reference to the broader framework of moral cognition and human–AI interaction developed in the preceding sections.

Chapter 7 synthesizes the central findings across all four studies. It discusses their theoretical implications for models of moral reasoning and advice-following, their practical relevance for the ethical design of AI systems, and their broader societal significance. The chapter concludes by outlining promising directions for future research on human–algorithm cooperation in moral and ethical domains.

2 Theoretical Background and Literature Review

The following chapter reviews the most relevant literature on morality and Artificial Intelligence. The first part examines how morality has been defined and conceptualized across philosophical and psychological traditions, tracing its historical and theoretical roots. I introduce the concept of intuitive moral with its neurological correlates. I elaborate how measuring morality works, how advice adds another layer in moral decision-making and finally, how advice utilization may be a heuristic to reduce cognitive effort. In the second part, I introduce Artificial Intelligence (AI), outline its development, and discuss its growing role in advisory contexts. The third part integrates both domains by examining how individuals perceive AI in moral decision-making situations and whether AI systems can themselves be considered moral agents. The chapter concludes with a review of empirical research on AI moral advice and its implications for human decision-making as well as the aim for the present dissertation.

2.1 Morality: Definitions and Historical Foundations

2.1.1 Introduction to Morality

Morality is a cornerstone of human social life, yet its meaning and scope remain widely debated across disciplines. People frequently invoke moral language to express judgments about right and wrong, fairness, or justice, but these expressions rest on different assumptions about what constitutes morality and where its authority originates. Some scholars regard morality as a system of universal principles related to harm avoidance, fairness, and human dignity (Scanlon, 2011; Turiel, 1985), whereas others emphasize its cultural and contextual variability, viewing moral codes as products of shared social norms and traditions (Haidt, 2008; Shweder et al., 2013). Despite these differences, most definitions converge on the idea that morality provides prescriptive standards for evaluating actions, intentions, and character. These are standards that claim a form of authority distinct from mere social convention (Gert

& Gert, 2002; Graham et al., 2011).

Philosophers have long debated the foundations of this authority and the nature of moral obligation. Early written sources, such as Mesopotamian law codes, the Hindu Vedas, the Egyptian Instructions of Amenemope, and the Hebrew Bible, reveal that moral prescriptions historically functioned to preserve social order and cooperation. The Code of Hammurabi formalized justice, reciprocity, and accountability as civic virtues, while the Hebrew Bible placed moral knowledge at the heart of human existence. In classical Greek philosophy, morality became a subject of systematic inquiry: Plato's Republic investigated the nature of justice and the good life, whereas Aristotle's Nicomachean Ethics described morality as the cultivation of virtue through habituation and practical wisdom (phronesis) (Aristoteles et al., 2011; Haidt, 2008; Haidt & Kesebir, 2010; Scanlon, 2011).

Enlightenment thinkers later redefined morality in terms of rationality and universality. David Hume (1739) argued that moral judgments arise from sentiment and sympathy rather than abstract reason. In contrast, Immanuel Kant (1998) grounded morality in pure practical reason, proposing the categorical imperative as an unconditional moral law. John Stuart Mill (1966) subsequently developed utilitarianism, locating moral worth in the consequences of actions and the maximization of collective happiness. These three frameworks, virtue ethics by Hume (1739), deontology by Kant (1998) and consequentialism by Mill (1966) remain the major reference points for contemporary moral philosophy and form the conceptual foundation for modern moral psychology.

From a psychological perspective, morality can be understood as a multi-level system encompassing moral reasoning, emotion, and behavior (Blasi, 1983; Ellemers et al., 2019). It regulates interpersonal relations, motivates prosocial behavior, and sustains social cohesion. While early developmental theorists such as Piaget (1934) and Kohlberg (1985) emphasized moral reasoning as the highest form of moral maturity, later intuitionist and dual-process approaches by J. D. Greene et al. (2004) and Haidt (2001) among others argue that moral judgment is primarily shaped by intuitive and affective processes (J. I. M. Carpendale, 2009; J. I. Carpendale, 2000; Haidt & Kesebir, 2010). Thus, contemporary moral psychology views morality as both a cognitive and an emotional phenomenon, rooted in human evolution, guided by social interaction, and dynamically responsive to cultural norms.

The following sections expand on these ideas by first outlining major ethical frameworks and psychological models of moral development and then introducing modern approaches that emphasize the role of intuition, emotion, and social context in moral cognition.

2.1.2 Major Ethical Frameworks

The evolution of moral inquiry from "Who must I become?" to "What is the right thing to do?" represents a pivotal shift in ethical thought. Two dominant frameworks rose from this change of thinking about morality. On the one hand, Deontology, articulated by Immanuel Kant (1998), evaluates morality based on adherence to universal duties and principles, independent of outcomes. Kant's principle of universalizability asserts that actions are moral if their guiding rules can consistently apply to all rational beings. Conversely, consequentialism, formulated by John Stuart Mill (1966), judges actions by their outcomes, advocating for those that maximize overall good or utility.

Psychological Approaches to Moral Development

Building on the philosophical foundations of deontology and consequentialism, psychological theories have enriched our understanding of morality by exploring how it develops across the human lifespan. Rather than focusing solely on abstract principles or the consequences of actions, these theories examine how individuals come to understand and apply moral reasoning as they grow and interact with others. Jean Piaget (1934) laid the groundwork for developmental approaches to morality by identifying how children's moral understanding evolves through distinct stages. In his studies, Piaget observed that young children initially view rules as fixed and externally imposed by authority figures (heteronomous morality). Over time, through cognitive development and social interactions, children transition to autonomous morality, where they begin to see rules as flexible agreements made between individuals. This shift reflects the growing ability to think abstractly, empathize with others, and engage in cooperation (J. I. M. Carpendale, 2009). His research provided the foundation for future studies of moral development, most notably the work of Lawrence Kohlberg (1985). Building on Piaget's insights, Kohlberg (1985) proposed a structured, stage-based model of moral reasoning. Kohlberg extended Piaget's concept of moral autonomy by emphasizing

the developmental progression from self-interest to universal ethical principles. His theory, which remains a cornerstone in the study of ethics and psychology, describes how individuals progress through six stages of moral reasoning, grouped into three levels: pre-conventional, conventional, and post-conventional.

At the pre-conventional level, morality is primarily self-centered, with decisions guided by the desire to avoid punishment (Stage 1) or gain rewards (Stage 2). For example, children at this level might follow rules solely because they fear consequences or expect personal benefits. As individuals move to the conventional level, their reasoning becomes shaped by societal norms and expectations. In Stage 3, moral behavior is motivated by the desire to maintain relationships and meet the expectations of close social groups, while Stage 4 emphasizes adherence to laws and social order for the greater stability of society. The highest stage of moral reasoning is found at the post-conventional level, where individuals base their decisions on universal ethical principles that transcend societal norms. In Stage 5, people recognize the importance of balancing individual rights with the collective good, understanding that laws are social constructs that can be questioned or amended. At Stage 6, individuals develop a principled conscience grounded in ideals such as justice, equality, and respect for human dignity. Kohlberg emphasized that these higher stages represent more morally adequate reasoning because they resolve ethical conflicts through universalizable and reversible principles ("justice as reversibility") (Kohlberg, 1985).

Critiques of Stage-Based Models

Efforts to create stage-based models of moral development, while foundational in psychology, have been criticized for their rigidity, cultural bias, limited scope, and neglect of emotional dimensions. These models, such as those proposed by Piaget and Kohlberg, often emphasize linear progressions and universal applicability, yet they may fail to account for the coexistence of different types of moral reasoning, as well as the influence of adult-child relationships alongside peer interactions. Furthermore, their Western-centric focus on autonomy and justice overlooks moral values such as relational harmony, which are emphasized in collectivist cultures. Additionally, these models rely heavily on abstract, hypothetical dilemmas, like Kohlberg's Heinz dilemma, which lack ecological validity and fail to capture the emotional

and situational complexities of real-world moral decision-making (J. I. M. Carpendale, 2009).

Key Thematic Areas in Modern Moral Research

Throughout history, moral reasoning has been examined by various bright minds across great societies (e.g., Kant (1998), Kohlberg (1985), and Mill (1966)). However, modern investigations often present an incomplete understanding, neglecting critical aspects of human morality articulated in influential theoretical frameworks. The prevailing challenge for the field is to account for the complexity and multifaceted nature of morality by integrating diverse psychological antecedents and implications and fostering connections between mechanisms rather than studying them in isolation (Ellemers et al., 2019). Key thematic areas include moral reasoning, moral judgments, moral behavior, moral emotions, and moral self-views, each addressing distinct aspects of how individuals and groups navigate ethical questions. Moral reasoning explores the principles and guidelines people use to determine what is right or wrong, often drawing on abstract ideals or situational considerations. Moral judgments evaluate the morality of others' actions or character, shaping social interactions and perceptions. Moral behavior examines the practical expression of these judgments and reasoning in actions, such as helping, cooperating, or engaging in deceit. Moral emotions, such as guilt, shame, or outrage, are integral to moral experiences, guiding behavior while reflecting deeply ingrained social and cultural norms. Moral self-views focus on how individuals perceive their own morality, balancing their self-image with their actions and often employing strategies like self-justification to maintain a sense of moral consistency. These thematic areas highlight the complexity of morality as it operates on individual, interpersonal, and group-based levels, while being shaped by both rational deliberation and emotional intuition. By integrating these domains, contemporary research seeks to provide a comprehensive understanding of morality and its implications for addressing pressing ethical questions in increasingly diverse and interconnected societies (Ellemers et al., 2019).

2.1.3 Defining Morality and Moral Ethics

The understanding of morality has evolved significantly over time, with each iteration refining and expanding upon prior conceptualizations. Early foundational work by Turiel (1985, p.3)

provided a precise framework for the moral domain, defining it as comprising "prescriptive judgments of justice, rights, and welfare pertaining to how people ought to relate to each other," emphasizing morality as a guide for interpersonal relations, focusing on justice and fairness as central elements. Building on this foundation, Gert and Gert (2002) introduced a dual perspective of morality: descriptive and normative. Descriptively, morality refers to specific codes of conduct proposed by societies, groups, or individuals, capturing its cultural and contextual variability. Normatively, morality entails a universal code of conduct that would be endorsed by all rational people under specified conditions, signaling a broader philosophical aim to identify universal moral principles. As the field of moral psychology matured, Haidt (2008) offered a transformative expansion of Turiel's framework. He incorporated functional and systemic elements into the definition, emphasizing the practical role of morality in human social life. According to Haidt, moral systems are designed to regulate selfishness and foster cooperation, facilitating harmonious social interactions and community cohesion. This marked a shift from a solely principle-based view to one recognizing the adaptive and integrative functions of morality in human evolution (Haidt, 2008). Further elaboration came from Haidt and Kesebir (2010, p.7), who extended this perspective by presenting morality as a dynamic and multidisciplinary field. They described moral systems as "interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms" that enable cooperative social life. This expanded definition highlighted the complexity of moral systems and their role in aligning individual behavior with collective well-being. They identified three key principles of modern moral psychology. First, they noted the intuitive primacy of morality, where fast, automatic moral intuitions often precede reasoning. Second, they argued that moral reasoning functions as a tool for navigating social interactions rather than uncovering objective truths. Third, they emphasized the role of moral binds, shared moral frameworks that foster group cohesion and collective identity. Their work also introduced virtue ethics into the discussion, framing virtues as cultivated traits deeply rooted in cultural contexts. Rather than adhering to abstract principles, virtue ethics emphasizes character development through practice, habit, and the influence of moral exemplars. Morality can shape perceptions and emotions over a lifetime, and enriches its applicability in diverse cultural and social settings.

Integrating Definitions

Based on the previous elaborations and approaches of defining morality, we aimed to integrate all into one comprehensive definition of morality that serves the purpose of this present research. Merging previous definitions into one, we go with the following definition:

> "*Morality is a dynamic system of values, norms, and psychological mechanisms that regulate individual behavior to promote cooperation and harmonious social interactions.*"

This definition reflects the evolving understanding of morality as a flexible system through different phases of moral research, adapting to human cooperation and social cohesion. By bridging ethical principles with real-world interactions, we believe it provides a comprehensive foundation for our aim of studying moral decision-making. The following chapter explores the most modern approaches to morality and ultimately aims to create a common thread that links different approaches and concepts. Therefore, we put a special emphasis on intuitive moral approaches.

2.1.4 Modern Approaches to Morality

Modern approaches to morality emphasize the interplay between intuition, reasoning, and evolutionary foundations in shaping moral judgments. Unlike earlier rationalist models, contemporary theories recognize morality as a product of both cognitive processes and social adaptation. The Social Intuitionist Model (SIM), Moral Foundations Theory (MFT), and the new synthesis of morality are interconnected frameworks that build upon one another to form a cohesive understanding of morality. Each step in their development refines and expands the previous one, creating a unified progression in moral psychology. The Social Intuitionist Model (SIM) laid the groundwork by emphasizing the primacy of intuition in moral judgment. Haidt (2001) challenged reasoning-centered models by proposing that moral decisions are largely driven by emotional, automatic intuitions, with reasoning serving primarily as a post-hoc justification. This intuitionist perspective introduced the idea that morality is deeply rooted in emotional processes and shaped by social interactions, setting the stage for further exploration into the structure and diversity of moral judgments. Building on SIM,

Moral Foundations Theory (MFT) expanded the scope by identifying specific domains or foundations (Care, Fairness, Loyalty, Authority, and Sanctity) of morality, which provided a framework for understanding how moral intuitions are shaped by evolutionary pressures and refined by cultural contexts. MFT extended SIM's focus on intuition by categorizing the moral domains that intuitions operate within, offering a more detailed understanding of how and why moral priorities differ across cultures and ideologies (Graham et al., 2013). The new synthesis of morality represents the culmination of these ideas, integrating insights from SIM and MFT into a broader, multidisciplinary framework. It situates the intuitive and foundational aspects of morality within an evolutionary and social context, emphasizing how morality functions not only at the individual level but also in maintaining social cohesion and managing intergroup dynamics. The new synthesis incorporates findings from neuroscience, psychology, and anthropology to explain morality as an adaptive system that binds individuals into cooperative groups while creating boundaries between them.

Before detailing these moral approaches, we briefly digress to outline dual-process theories of cognition. This detour provides the general cognitive architecture (automatic, intuitive processing versus controlled, reflective processing) that underlies contemporary intuitionist accounts. We then return to SIM and MFT, showing how they instantiate this architecture in moral cognition and how the New Synthesis situates these mechanisms within broader social and evolutionary functions.

Dual-Process Theories of Moral Cognition

Dual-process theories describe human cognition as the product of two qualitatively distinct but interacting modes of processing: one that is fast, automatic, associative, and affective, and another that is slow, controlled, and rule-based (J. Evans & Frankish, 2009; J. S. B. T. Evans, 2008, 2003; Kahneman, 2011). Often described as Type 1 and Type 2 processing, these modes capture differences in processing style rather than anatomically separable cognitive systems (Keren & Schul, 2009; Melnikoff & Bargh, 2018). Type 1 processing operates intuitively and effortlessly: it draws on learned associations, pattern recognition, and affective cues to generate rapid judgments with minimal working-memory demand. It is context-sensitive, parallel, and evolutionarily older, enabling adaptive responses in environments that require

quick appraisal or action. In contrast, Type 2 processing engages deliberate, sequential, and rule-based reasoning. It depends on cognitive resources such as attention and working memory, and is therefore slower, more flexible, and better suited for abstract reasoning, hypothetical thinking, and the evaluation of novel problems (J. S. B. T. Evans & Stanovich, 2013; J. S. Evans, 2003).

The classic heuristics-and-biases tradition by Tversky and Kahneman (1973, 1974) demonstrated that Type 1 processing relies on heuristics, efficient mental shortcuts that simplify judgment but can systematically bias reasoning. These heuristics operate through a process of attribute substitution: when a complex target attribute is difficult to assess, it is unconsciously replaced by a simpler, more accessible cue (Frederick, 2005; Kahneman & Frederick, 2002). In moral contexts, such substitutions may involve intuitive cues like perceived harm, purity, or authority endorsement, producing swift yet occasionally inconsistent moral intuitions (Sunstein, 2005). We further elaborate moral heuristics in chapter 2.1.6.

Building on these findings, J. Evans and Frankish (2009) and J. S. B. T. Evans (2008) formalized two interacting dual-process architectures: a default–interventionist model, in which Type 1 processing delivers an initial intuitive response that Type 2 processing may monitor and override, and a parallel–competitive model, in which both types of processing operate simultaneously and compete for behavioral control. In both views, intuitive processing typically dominates under time pressure, cognitive load, or emotional arousal, while reflective reasoning emerges when motivation and resources permit deliberate correction. However, critics caution that "two systems" language may overstate ontological separation, arguing that dual-process distinctions are better conceived as differences in processing mode or resource demand rather than as discrete cognitive systems (Keren & Schul, 2009; Melnikoff & Bargh, 2018).

In moral psychology, these dual-process mechanisms provide the cognitive foundation for intuitionist theories: the Social Intuitionist Model (SIM) posits that moral judgments are primarily intuitive and socially modulated, whereas the Moral Foundations Theory (MFT) identifies recurrent intuitive domains that structure these judgments across cultures. These approaches translate the general dual-process architecture into models of moral cognition that link affective intuition and reflective reasoning in the production of moral judgment.

Social Intuitionist Model

Haidt (2001) proposed the Social Intuitionist Model (SIM) to explain moral judgment as a process predominantly guided by intuitive, emotional responses rather than deliberate reasoning. To illustrate this, Haidt uses the provocative example of adult siblings, Julie and Mark, who consensually engage in a sexual act during a vacation. The scenario is carefully designed to remove harm or negative consequences: they use contraception, mutually agree on the act, and ensure secrecy. Despite the absence of harm or objective consequences, most individuals instinctively judge their behavior as morally wrong. Haidt highlights that this reaction stems from immediate moral intuitions, not reasoned deliberation. When pressed to justify their judgment, people often struggle to articulate a coherent rationale, instead defaulting to vague references to social taboos or norms. This phenomenon is termed moral dumbfounding and underscores a key principle of the SIM: moral reasoning primarily functions as a post-hoc rationalization of intuitive judgments, rather than as the driver of those judgments (Haidt & Joseph, 2004). The example of Julie and Mark demonstrates the powerful influence of evolutionary and cultural factors on moral intuitions. Deep-seated taboos against incest and visceral feelings of disgust serve as emotionally charged mechanisms that guide moral judgment, independent of logical considerations about consent or harm. This claim is supported by neuroimaging research by J. D. Greene et al. (2004), which demonstrates that moral decision-making engages brain regions associated with emotional processing. Their findings suggest that intuitive, affective responses play a dominant role in moral cognition, with deliberate reasoning emerging as a secondary process primarily used to rationalize or justify intuitively formed judgments. We will elaborate on these findings in more detail in chapter 2.1.5. This highlights the primacy of intuition in moral cognition, situating deliberate reasoning as a secondary process used to explain or justify the conclusions already formed by intuition. The SIM further emphasizes the social nature of morality. Haidt (2001) and Haidt and Joseph (2004) argue that moral reasoning often functions as a tool for social negotiation, rather than solitary reflection. Individuals use reasoning to persuade others, defend their judgments, or build alliances within social groups. Through social interactions, feedback from peers can refine or reshape an individual's moral intuitions, creating a dynamic interplay between personal cognition and cultural norms. This feedback loop ensures that morality

is continually influenced by both interpersonal and societal contexts. In redefining moral cognition, the SIM suggests that morality is not a product of purely rational processes but emerges from the interaction of intuition, emotion, and social dynamics. This challenges the traditional view of morality as driven by conscious reasoning and positions emotional intuition as the central mechanism in moral judgment.

Moral Foundations Theory

Moral Foundations Theory (MFT), initially proposed by Haidt and Joseph (2004) and further developed by Graham et al. (2011, 2013, 2016) and Haidt (2012), provides a framework for understanding the psychological mechanisms underlying moral judgments that was designed to explain both the variations and common aspects of moral judgments across cultures. Building upon the Social Intuitionist Theory (Haidt, 2001), MFT extends the idea of intuition over reasoning in moral decision-making by systematically categorizing moral intuitions into distinct, evolutionarily shaped foundations. Unlike SIT, which broadly described morality as intuition-based, MFT identifies specific domains of moral concern, each of which evolved to address recurrent social challenges in human history. Furthermore, MFT provides an explanatory model for why different cultures and political ideologies prioritize different moral values, a refinement beyond SIT's focus on individual moral cognition.

According to the Moral Foundations Theory (Graham et al., 2011, 2013; Haidt, 2012; Haidt & Joseph, 2004), morality is rooted in a set of evolutionarily adapted psychological systems, referred to as moral foundations, which respond to distinct social challenges. These foundations are thought to represent the innate "first draft" of the moral mind, which is subsequently shaped and refined by cultural influences. Initially, Haidt and Joseph (2004) proposed five primary moral foundations: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation, which were later completed with the sixth foundation Liberty/Oppression (Haidt, 2012). These foundations are stable across WEIRD and Non-WEIRD samples (Doğruyol et al., 2019). Each foundation is hypothesized to emerge from recurrent adaptive challenges faced by humans over evolutionary time. For instance, the Care/Harm foundation is tied to the nurturing of vulnerable offspring, evoking emotions such as compassion or anger at perceived harm. Similarly, the Fairness/Cheating foundation reflects

sensitivities to reciprocal altruism, fostering trust and cooperation within social groups. The other foundations address challenges of group cohesion, hierarchy, and pathogen avoidance, respectively.

A key insight of MFT is its pluralistic approach to morality, which contrasts with traditional monistic theories that emphasize a single moral principle, such as justice or care. Instead, MFT argues that morality encompasses diverse and context-dependent values. This pluralism allows the theory to account for cross-cultural variation in moral emphasis (Graham et al., 2016). For example, Western societies often prioritize Care and Fairness, whereas collectivist cultures may place greater importance on Loyalty, Authority, and Sanctity.

MFT also underscores the role of intuition in moral cognition, building up and aligning with Haidt (2001)'s Social Intuitionist Model. Moral judgments, according to MFT, arise rapidly and emotionally, shaped by evolved intuitions. Deliberative reasoning often occurs post-hoc, serving to justify these intuitive responses rather than forming their basis. The theory highlights how cultural learning shapes the triggers for moral intuitions. For instance, the Sanctity foundation, originally evolved to avoid pathogens, may now be triggered by modern norms around dietary purity or sexual behavior.

The practical utility of MFT lies in its ability to bridge psychological, cultural, and political perspectives on morality. For instance, it explains ideological divides, with liberals emphasizing Care and Fairness, while conservatives value a broader set of foundations, including Loyalty, Authority, and Sanctity (Graham et al., 2009; Haidt & Graham, 2007). This integrative approach has been instrumental in advancing understanding of moral diversity and conflict across groups. MFT presents morality as a dynamic interplay between innate psychological systems and cultural influences. By emphasizing pluralism, intuition, and evolutionary roots, MFT offers a robust framework for studying how moral values arise and vary across individuals, cultures, and societies.

New Synthesis of Morality

The new synthesis describes a more interdisciplinary approach to explain morality as primarily driven by evolved emotional intuitions rather than conscious reasoning (Haidt, 2007, 2008, 2012; Haidt & Kesebir, 2010). It logically extends Moral Foundations Theory (MFT) by

integrating insights from neuroscience, cultural evolution, and group-level selection into a unified framework. While MFT established that moral judgments arise from distinct, evolutionarily shaped foundations, the New Synthesis explains how morality functions beyond individual cognition, serving social cohesion and cultural adaptation. It synergizes insights from social psychology, social-cognitive neuroscience, and evolutionary science, while incorporating perspectives from cognitive science, developmental psychology, economics, and philosophy (Haidt & Kesebir, 2010). By embedding moral intuitions in affective neuroscience, emphasizing the social function of moral reasoning, and highlighting morality's role in group dynamics, the New Synthesis refines and expands MFT. Rather than replacing it, this framework broadens the scope of moral psychology, offering a deeper understanding of both universal moral mechanisms and cultural variation. Modern Morality emphasizes three principles: intuitive primacy, morality's social function, and its role in binding communities (Haidt, 2007; Haidt & Kesebir, 2010).

(I) Intuitive Primacy The first principle of Intuitive Primacy asserts that moral judgments arise primarily from automatic, emotionally driven intuitions rather than from deliberate reasoning. While reasoning can refine, justify, or sometimes override these judgments, it typically follows intuition rather than generating moral conclusions independently. However, intuition is not absolute. In certain situations, reasoning can exert control over initial moral intuitions, particularly in cases of moral conflict or extended social deliberation (Haidt, 2007). The Principle of Intuitive Primacy builds on Haidt's Social Intuitionist Model (SIM) (Haidt, 2001), which posits that moral judgments arise primarily from intuitive processes rather than deliberate reasoning. The New Synthesis in Moral Psychology incorporates SIM as its foundation by strengthening the argument that moral cognition is fundamentally an intuitive process, shaped by emotions and social influences. Hauser et al. (2007) supports this view, arguing that moral intuition is an evolved cognitive adaptation, with an innate structure that guides human moral evaluations even in the absence of explicit reasoning or learned moral principles. Neuroscientific research, particularly by J. Greene and Haidt (2002), J. D. Greene et al. (2001, 2004, 2008, 2009), and J. D. Greene (2014), provides strong empirical support for the Principle of Intuitive Primacy by demonstrating how emotional and cognitive processes interact in shaping moral judgments. Their dual-process theory suggests that moral

judgments are primarily driven by automatic, emotionally charged intuitions, with controlled cognitive reasoning playing a secondary role. Results from neuroimaging studies reveal that moral judgments activate the amygdala, medial prefrontal cortex (mPFC), and posterior cingulate cortex (PCC), which are associated with emotional processing, before engaging cognitive control regions such as the dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC), which are involved in deliberative reasoning (J. D. Greene et al., 2004). This supports the claim that moral cognition is primarily intuitive rather than deliberative. Deontological judgments remain stable under cognitive strain, while utilitarian judgments suffer, indicating that moral intuition operates automatically, while moral reasoning requires effortful processing mediated by the DLPFC and ACC (J. D. Greene et al., 2008). People are significantly less likely to approve of direct harm, even when it results in greater overall benefits, highlighting the dominance of emotional aversion in moral decision-making, a response driven by the amygdala and ventromedial prefrontal cortex (vmPFC) (J. D. Greene et al., 2009). Dual-process models of moral cognition support the claim that intuitive, affective responses, mediated by the mPFC and amygdala, form the basis of deontological moral thinking, whereas utilitarian reasoning relies on the DLPFC to override these automatic responses in a slower, more deliberative manner (J. D. Greene, 2014).

(II) Moral Thinking is for Social Doing The second principle of the new synthesis of moral psychology is "Moral Thinking is for Social Doing." This principle suggests that moral reasoning is primarily a tool for social interaction rather than an objective search for truth. Haidt (2007) argues that moral reasoning functions more like that of a lawyer or politician, aiming to justify actions, maintain reputation, and persuade others, rather than serving as an impartial, truth-seeking mechanism. It aligns with an evolutionary and social-intuitionist perspective on moral cognition, emphasizing that human morality evolved to facilitate group cohesion, cooperation, and social influence (Haidt, 2007; Haidt & Kesebir, 2010). Haidt and Kesebir (2010) argue that moral reasoning is often post-hoc, meaning that it serves to justify actions and persuade others rather than to discover objective moral truths, again aligning with Haidt's Social Intuitionist Model, which suggests that moral judgments are primarily driven by intuitive processes, with reasoning playing a secondary role as a justification mechanism (Haidt, 2001). People use moral reasoning to advocate for their own interests or the interests of their

group rather than to seek moral truth. Studies indicate that individuals engage in motivated reasoning, selectively endorsing evidence that supports their pre-existing moral intuitions while dismissing contradictory information (Ditto et al., 2009). This pattern suggests that moral discourse functions similarly to legal or political argumentation, where individuals construct persuasive narratives rather than objective analyses. Moral reasoning also serves important functions in social life, particularly in regulating behavior through gossip and reputation management. Beyond individual cognition, moral discourse plays a key role in sustaining cooperation by reinforcing group norms and ostracizing those who violate them (Haidt & Kesebir, 2010). This is consistent with research by Fehr and Gächter (2002), who demonstrated that moral emotions and social norms are crucial for enforcing cooperation in groups.

(III) Morality Binds and Builds The third principle of moral psychology, "Morality Binds and Builds," posits that morality is not just about deciding what is right or wrong on an individual level. Instead, it plays a crucial role in strengthening social cohesion, facilitating cooperation, and creating large, stable groups. Moral values and shared beliefs act like social glue, ensuring the survival and prosperity of groups rather than merely serving individual ethical deliberation. Group selection plays a crucial role in shaping human moral behavior, as human morality evolved to promote cooperation within groups while also enabling competition between them. Empirical studies on cooperation and prosocial behavior provide additional support for the Morality Binds and Builds principle. Moral emotions such as guilt and outrage play a crucial role in enforcing cooperative norms through altruistic punishment (Fehr & Gächter, 2002). Individuals are willing to incur personal costs to punish those who violate moral expectations, even in anonymous settings. The third principle is supported by neuroscientific research demonstrating that moral cognition is closely linked to brain regions responsible for social processing, empathy, and group cohesion. Moral decision-making consistently engages brain regions involved in social cognition, particularly the medial prefrontal cortex and the anterior cingulate cortex, supporting the idea that moral thought is fundamentally tied to social interaction rather than abstract reasoning (J. D. Greene et al., 2001; Lamm et al., 2011). These regions are also activated during empathy, indicating that moral cognition and social bonding share a common neural basis (Decety & Lamm, 2006; Lamm et al., 2011). Empathy

for others' pain elicits activity in the anterior insula and the medial cingulate cortex, regions also engaged in the direct experience of pain, suggesting that moral behavior relies on neural mechanisms designed for social bonding and cooperation (Decety & Lamm, 2006). Individuals with stronger neural responses in these areas show greater prosocial tendencies, reinforcing the role of empathy in moral action (Decety & Lamm, 2006). Different experimental paradigms reveal distinct activation patterns, with direct visual representations of suffering engaging motor-related areas linked to action understanding, while abstract cues of pain activate brain regions associated with mentalizing and perspective-taking, highlighting the flexibility of moral cognition in social contexts (Lamm et al., 2011). These findings collectively support the idea that morality evolved as a mechanism for fostering group cohesion and regulating social behavior through shared emotional and cognitive processes (Decety & Lamm, 2006; Lamm et al., 2011). Morality functions primarily as a social mechanism, reinforcing cooperation, trust, and group cohesion rather than serving purely individual ethical reasoning.

Yet, the New Synthesis in Morality is not without criticism. It has been criticized for overemphasizing intuition while underestimating the role of rational deliberation in moral development, personal responsibility, and ethical decision-making (Kristjánsson, 2016). While moral intuitions are fundamental, reasoning also plays a crucial role in shaping moral growth and navigating complex ethical dilemmas. Additionally, research on cooperation and social norms highlights the importance of deliberative processes in maintaining moral order within groups (Fehr & Gächter, 2002). Overall, the New Synthesis in moral psychology presents morality as an evolved mechanism for fostering cooperation, social cohesion, and group stability. It emphasizes intuitive moral judgments over rational deliberation, framing moral reasoning as a social tool rather than an impartial search for truth. While some argue it underestimates deliberation, this perspective offers a comprehensive framework for understanding moral behavior across cultures and its role in shaping human societies.

Synthesis of Intuitive Moral Theories

Modern morality is a dynamic and evolving construct that reflects the interplay between human intuition, social structures, and cultural influences. It is not a rigid set of absolute principles but rather a fluid system that adapts to societal changes, technological advancements, and global

challenges. At its core, morality functions as a mechanism for regulating behavior, fostering cooperation, and maintaining social harmony. It arises from deeply ingrained emotional responses and cognitive processes that shape individual and collective decision-making. Moral cognition is guided by an intrinsic sense of fairness, empathy, and responsibility, which help individuals navigate social interactions and reinforce ethical conduct. However, morality is not merely an individual pursuit; it is deeply embedded within cultural traditions and institutional frameworks that shape moral norms and expectations. As societies become more interconnected, moral systems need to expand beyond parochial loyalties and integrate broader ethical considerations that address collective well-being and sustainability. A comprehensive understanding of morality acknowledges the balance between personal values and collective responsibilities. While individual morality is often shaped by intuition and immediate social contexts, societal morality extends to norms that promote justice, equality, and mutual respect. This interplay between individual and societal morality ensures that ethical principles remain relevant and adaptable to contemporary challenges. The evolution of morality also reflects the growing recognition of shared human experiences and universal values that transcend cultural boundaries. Issues such as environmental stewardship, human rights, and technological ethics require a cooperative and inclusive approach to morality, that prioritizes global responsibility over insular traditions. In this light, morality is not a static doctrine but a continually refined system that aligns with both human nature and societal progress.

Moral decision-making is often shaped by rapid, instinctive reactions rather than deliberate, rational thought. Many moral choices occur in emotionally charged situations, where individuals respond intuitively rather than carefully analyzing all possible consequences. This intuitive nature of moral judgment suggests the involvement of dual cognitive processes, making Dual Process Theory (DPT) a highly relevant framework for understanding moral decision-making.

2.1.5 Measuring Morality

Defining morality provides a theoretical framework for understanding what constitutes ethical behavior, but practical inquiry into how individuals make moral decisions requires tools and methods that can capture the nuances of these processes. Measuring morality transforms abstract concepts into empirical data, enabling researchers to examine not only what people

believe to be right or wrong but also how they navigate complex ethical scenarios, shifting to a more scientific approach. Moral dilemmas provide an effective and insightful tool to measure morality because they present individuals with structured scenarios that require navigating conflicting ethical principles. They present individuals with situations involving conflicting ethical principles and act as structured scenarios that force difficult trade-offs, revealing the guiding principles that underpin moral reasoning (Christensen & Gomila, 2012). In this chapter, we examine the quantification and empirical assessment of moral behavior, exploring both its theoretical foundations and practical applications. We analyze moral dilemmas as the key methodological approach, assessing their strengths and limitations in capturing moral decision-making processes. Furthermore, we discuss the advantages and limitations whilst incorporating perspectives on everyday moral choices and distinguishing between utilitarian and deontological reasoning, as well as altruistic and egoistic motivations. Finally, we consider the neurobiological underpinnings of moral decision-making, highlighting the cognitive and affective mechanisms that drive ethical judgments.

Theoretical and Practical Implications for Measuring Morality

Theoretical Implications Measuring morality provides insights into the cognitive and emotional processes that underlie moral judgments, allowing researchers to refine and develop theoretical models of moral reasoning. Theories like the Dual-Process Theory of Moral Judgment (J. D. Greene et al., 2001) and Moral Foundations Theory (Haidt & Graham, 2007) rely on empirical data to explain how individuals navigate ethical dilemmas. These models distinguish between intuitive, emotion-driven moral judgments and those based on deliberate reasoning, furthering the understanding of how moral cognition operates in different contexts and providing valuable insights into human moral decision-making. By systematically assessing moral decision-making, researchers can clarify the distinction between deontological ethics (rule-based morality) and consequentialist ethics (outcome-based morality), as well as differentiate between altruistic and egoistic decisions. This helps in identifying which principles individuals apply in specific contexts, providing deeper insights into the fundamental mechanisms guiding moral reasoning. Additionally, comparative studies using moral dilemmas reveal whether moral judgments are universal or culturally dependent. Cross-cultural research

suggests that certain moral intuitions, such as aversion to harm, may be universal, whereas values like fairness or loyalty may vary across cultures (Bago et al., 2022). Measuring morality advances interdisciplinary research in moral psychology, neuroscience, and philosophy. It enables researchers to investigate the neurological basis of moral reasoning and the role of emotions in ethical decision-making (J. D. Greene et al., 2004, 2008, 2009). These insights contribute to a more comprehensive understanding of the interplay between affective and cognitive processes in moral cognition. The empirical study of morality bridges the gap between abstract philosophical theories and experimental psychology, allowing for practical validation and refinement of ethical concepts. Philosophical theories of morality often rely on rationalist assumptions, but empirical studies using moral dilemmas allow scientists to test these theories in real-world contexts, reinforcing or challenging long-held assumptions about human moral behavior.

Practical Implications Beyond theoretical advancements, the ability to measure morality has profound practical implications. Understanding moral reasoning is critical for addressing key societal challenges in law, policy-making, and technological ethics. For instance, in medical ethics, empirical research on moral decision-making informs policies on end-of-life care, healthcare resource allocation, and genetic engineering, where difficult moral trade-offs are required (Singer et al., 2019). Ethical dilemmas in medicine, such as prioritizing patients in critical care or evaluating the moral permissibility of genetic modifications, necessitate a rigorous understanding of how moral judgments are formed.

Morality measurement is also essential in the development of ethical AI and autonomous systems. The integration of moral reasoning into self-driving cars, medical decision-making algorithms, and automated legal frameworks depends on empirical findings about how humans make ethical choices (Bonnefon et al., 2016). By incorporating moral principles into AI decision-making, researchers and engineers can create systems that align more closely with human ethical intuitions, minimizing risks associated with automated ethical reasoning. Identifying systematic moral biases, such as in-group favoritism or implicit moral hypocrisy, allows institutions to develop policies that promote fairness, equity, and moral integrity in governance and organizational ethics (Bauman et al., 2014). Understanding these biases can inform policies on discrimination, corporate responsibility, and institutional accountability,

ensuring that decision-making processes reflect broader ethical considerations rather than individual or group-based biases. As moral and ethical considerations become increasingly relevant in a rapidly evolving world, it is crucial to integrate moral psychology insights into education, leadership training, and public policy. The continuous growth of technological and societal complexities underscores the necessity of empirical moral research, ensuring that ethical standards remain robust and adaptable. By advancing the scientific study of moral decision-making, researchers can contribute to a more informed and ethically responsible society that is equipped to navigate the moral challenges of the future.

Moral Dilemmas and Measures

The study of morality often relies on structured methods to assess how individuals navigate ethical decision-making. Among these, moral dilemmas serve as a crucial tool in psychological research, allowing scholars to examine the cognitive and emotional mechanisms underlying moral judgments.

A moral dilemma, as defined by Duff and Sinnott-Armstrong (1989), occurs when an individual faces a conflict between two or more non-overridden moral requirements, meaning that neither moral obligation can be dismissed in favor of the other. In such cases, the individual is morally required to act in two incompatible ways, yet neither obligation can be considered stronger or more justifiable than the other. This creates an unresolvable ethical conflict, where any decision leaves a moral residue, such as guilt or regret, regardless of the chosen action.

These dilemmas present individuals with scenarios that force them to weigh competing moral principles, often revealing deep-seated ethical intuitions and reasoning processes, ultimately forcing decisions between two main conflicting moral principles (Bago et al., 2022). Commonly, a moral dilemma is a short story about a situation involving a moral conflict. A moral conflict is a situation in which the subject is pulled in contrary directions by rival moral reasons. It entails the awareness of the incompatibility of two courses of action and their subsequent outcomes.

Moral conflicts can be of many different types, such as (i) conflicts between personal interests and accepted moral values, (ii) conflicts between different duties, (iii) conflicts between sets of

apparently incommensurable values, and even (iv) conflicts stemming from one unique moral principle. Typically, the scenario that we face in a moral conflict or dilemma is that both options have important moral reasons to support them (Christensen & Gomila, 2012).

We regard moral dilemmas as a structured experimental stimulus consisting of specific design parameters and independent variables. In this context, moral dilemmas function similarly to theoretical or statistical models, where researchers can manipulate different variables to assess their influence on moral decision-making. By systematically adding or removing parameters, it becomes possible to identify which elements play a crucial role in shaping moral judgments and behavioral outcomes (Hauser et al., 2007). These elements must be systematically identified and controlled in all empirical studies investigating moral judgment.

The Trolley Dilemma: A Famous Example

The classic trolley problem is a well-established example of a moral dilemma first introduced by Philippa Foot (1967). The dilemma presents a scenario where a runaway tram is heading down a track where five workers are present and will be killed if the tram continues its course. The driver of the tram has the option to steer the vehicle onto another track, where only one worker is present. This action would inevitably cause the death of the one worker but would save the five. There are two possible outcomes that arise from this dilemma, both including harm to fictitious workers. Making a decision that avoids harm entirely is not an option in this scenario, as both outcomes involve the loss of life. This leaves you with two distinct decisions that both include the harm of one or more individuals. Your decision is based on the moral framework you apply to the decision, with no objectively correct or universally accepted answer. Either you pull the lever to minimize harm and save the greatest number of lives, or you choose not to pull the lever, thereby refraining from taking action but accepting the inevitable death of five instead of one, regardless of the net benefit (Foot, 1967). Both moral frameworks are discussed in further detail in the following passage.

The perception of moral dilemmas can change significantly when the parameters of the scenario are altered, revealing the complexity and context-dependence of moral judgment. For instance, in the trolley problem, imagine that the single person on the sidetrack is a convicted criminal, while the five individuals on the main track are children or doctors. Such variations

challenge the initial ethical calculus and highlight how factors like the perceived social value, innocence, or personal connection to those involved influence moral decision-making. Similarly, questions arise when considering if the victim is a close relative, a foreigner, or someone with an essential societal role (Christensen & Gomila, 2012). The trolley dilemma is designed to explore how individuals navigate moral conflicts and make ethical decisions when faced with competing principles. The scenario presents a clear choice: actively intervene to save five people at the cost of one life or refrain from action, allowing the greater loss of life to occur. This dilemma highlights the distinction between actively causing harm and passively allowing harm, a central principle in deontological ethics.

There are various types of moral dilemmas that illustrate the distinction between utilitarian and deontological decision-making, as explored in studies such as J. D. Greene et al. (2008) and others. They all share the distinction between utilitarian and deontological choices, where, from a consequentialist perspective, minimizing harm and maximizing overall well-being is the preferred choice, as it aligns with the principle of achieving the greatest good for the greatest number. In the case of the trolley dilemma, this would justify pulling the lever to divert the trolley, as it results in the least overall harm. In contrast, a deontological perspective holds that actively causing the death of an innocent person constitutes a fundamental moral violation, regardless of the consequences. This brings us to two main conflicting key principles in ethics: the decision between utilitarian and deontological choices.

Deontological and Utilitarian Choices

The distinction between utilitarian and deontological choices is central to moral psychology, as it reveals the competing cognitive and emotional mechanisms that shape moral decision-making.

A utilitarian choice is based on outcome-based morality, where the ethical acceptability of an action is determined by its consequences. Utilitarian (also referred to as consequentialist) philosophies hold that an action is morally acceptable if it maximizes well-being for the greatest number of people (in terms of saved lives, for example) (Mill, 1966). In the context of moral dilemmas, a utilitarian response favors actions that lead to the greatest benefit for the greatest number, even if they involve harm to an individual. This is evident in sacrificial

dilemmas like the trolley problem, where pulling the lever to sacrifice one person in order to save five aligns with utilitarian principles (Gawronski & Beer, 2016). Utilitarian judgments are often associated with deliberate cognitive reasoning, involving the calculation of costs and benefits rather than relying on emotional responses (J. D. Greene et al., 2004).

A deontological choice, by contrast, is based on rule-based morality, where the moral permissibility of an action depends on its adherence to moral norms rather than its consequences. This perspective is most closely associated with Immanuel Kant (Kant, 1998), who argued that moral actions must conform to categorical imperatives (universal moral laws that apply regardless of situational outcomes). From a deontological standpoint, an action is morally unacceptable if it violates fundamental moral rules, such as the prohibition against killing, even if breaking the rule would lead to a better overall outcome (Gawronski & Beer, 2016). Deontological judgments are typically linked to intuitive and emotional processes, as individuals experience a moral aversion to directly harming others, even when rational cost-benefit analyses suggest doing so would lead to a better outcome (J. D. Greene et al., 2001).

Deontological and Utilitarian Choices in the Trolley Dilemma

Utilitarian Perspective Translated to the trolley dilemma, the choice of action differs fundamentally depending on whether a person adopts a utilitarian or deontological perspective. From a utilitarian standpoint, the morally correct decision is to pull the lever, diverting the trolley onto the sidetrack where it will kill one person instead of five. This choice is justified by the principle of maximizing overall well-being, meaning that sacrificing one life to save five results in the greatest net benefit. According to the utilitarian point of view, moral actions should be evaluated based on their consequences, using a cost-benefit analysis to determine the morally optimal outcome. That is, when people engage in rational reasoning and suppress their emotional aversions, they are more likely to justify pulling the lever as a necessary action to minimize overall harm (Gawronski & Beer, 2016).

Deontological Perspective In contrast, a deontological approach dictates that the individual should not pull the lever, as actively diverting the trolley directly causes the death of

an innocent person, violating the fundamental moral rule against killing. According to deontological ethics, moral actions must be guided by categorical imperatives, meaning that certain moral rules, such as "do not kill", are absolute and cannot be broken, regardless of the consequences. From this perspective, pulling the lever would be morally impermissible because it involves actively choosing to harm an innocent person, even if it leads to a greater good. This contrasts with the notion of omission, where allowing the trolley to continue its path is not seen as morally equivalent to directly intervening and causing harm (Gawronski & Beer, 2016).

Advantages and Disadvantages in the Usage of Dilemmas as a Measure of Moral Behavior

Advantages Moral dilemmas offer several key advantages in research (Christensen & Gomila, 2012). Moral dilemmas serve as a controlled experimental paradigm, allowing for the systematic manipulation of variables to analyze how distinct parameters influence moral judgment. This level of experimental control facilitates reliable comparisons across studies and ensures reproducibility in research.

By presenting conflicting moral principles, dilemmas provide an empirical method for distinguishing between utilitarian and deontological reasoning, enabling researchers to investigate the cognitive mechanisms underlying moral decision-making. Unlike simpler moral judgment tasks, dilemmas incorporate multiple variables, such as intent, emotional salience, and social context, offering a comprehensive analysis of moral cognition. Their structured nature allows researchers to isolate key factors that shape moral decision-making, such as intentionality, directness of harm, and consequences. This enables the investigation of framing effects, cognitive biases, and the role of emotions in ethical reasoning.

Additionally, the flexibility of moral dilemmas allows for variations in framing, wording, and presentation order to assess their impact on moral choices. Moral dilemmas are widely used in cross-cultural studies to examine universal versus culturally specific moral intuitions, providing insight into how ethical principles vary across societies (Christensen & Gomila, 2012). They are also relevant for applied ethics, including policy decisions in medicine, law, and Artificial Intelligence, such as healthcare resource allocation and ethical programming of autonomous systems. Their ability to elicit controlled moral conflicts while maintaining

experimental precision makes them a valuable tool in modern research.

Disadvantages Yet, despite their widespread application and utility in moral research, moral dilemmas are not without methodological limitations and conceptual challenges and face criticism regarding their applicability to real-life decision-making. While hypothetical scenarios can predict cognitive and affective responses, they may not accurately capture the complexities of real-world ethical choices. Some argue that such scenarios lack experimental, mundane, and psychological realism, leading to concerns about their external validity (Bonnefon et al., 2019; FeldmanHall et al., 2012) and prioritize attention-grabbing over real-world applicability.

The dilemmas often reduce nuanced ethical conflicts into binary choices, which may oversimplify the complexities of moral reasoning in real-life situations (Christensen & Gomila, 2012). These dilemmas fail to represent real-world moral situations and may engage different psychological and neural mechanisms compared to realistic dilemmas (Himmelreich, 2018; Nyholm & Smids, 2016; Y. Zhang et al., 2023). They are too artificial and decontextualized, which limits their use for real-life ethics (Bauman et al., 2014; Gold et al., 2014). External validity of the measurement is an especially important factor in moral contexts, as it indicates the degree of reliability of the measurement. Here, Bauman et al. (2014) suggest that there are good reasons to worry that "judgment and decision-making processes people use in these unusual situations may not accurately reflect moral functioning in a broader set of situations" (pp. 536–537).

Specifically, three main objections have been raised: (i) these dilemmas are often perceived as intellectually amusing rather than ethically sobering, (ii) they fail to reflect the nuanced moral situations encountered in everyday life, and (iii) they do not elicit the same psychological processes as real-world moral decisions (Bauman et al., 2014). A significant challenge lies in the discrepancy between hypothetical and real-life moral choices. Studies indicate that individuals who endorse consequentialist choices in theoretical dilemmas do not necessarily act according to the same ethical principles in real-life situations (Bostyn et al., 2018; Sommer et al., 2010). This suggests that while moral dilemmas offer important insights into moral reasoning, they do not always serve as reliable predictors of real-world behavior. However, exposure to hypothetical scenarios can also shape real-life moral reasoning, highlighting a complex interaction between abstract ethical judgment and practical decision-making (Bostyn

et al., 2018).

A Solution: Everyday Moral Decision-Making A suitable solution for the lack of external validity might be real-life moral dilemmas, as they have been developed to better capture everyday decision-making (Singer et al., 2019; Sommer et al., 2010; Starcke et al., 2011). These vignettes involve decisions between moral obligations (e.g., helping others) and self-interest, avoiding extreme harm or legal consequences. Real-life scenarios provide a pathway to overcome the limitations of abstract moral dilemmas by reflecting the complexity of everyday ethical conflicts. They provide some key advantages compared to established high-conflict moral dilemmas. Unlike traditional sacrificial dilemmas, which often present extreme, life-or-death situations, everyday moral dilemmas are designed to reflect real-world ethical conflicts, enhancing their ecological validity. These dilemmas capture moral decisions that people encounter regularly, such as balancing personal interests with social obligations, thereby making them more applicable to daily moral reasoning. A key strength of everyday moral dilemmas is their flexibility in experimental design, allowing researchers to manipulate contextual variables such as social closeness (e.g., decisions involving family vs. strangers) and situational framing. This adaptability makes them useful for controlled studies in moral psychology while maintaining their real-world relevance. Moreover, Singer et al. (2019) addressed a methodological limitation of traditional moral dilemmas by developing two parallelized item sets, enabling within-subject experimental designs while minimizing memory effects and interindividual variability. Compared to high-conflict dilemmas, which often lead to extreme binary responses (e.g., choosing between saving one life or five), everyday dilemmas provide greater response variability, reducing ceiling and floor effects in moral decision-making research. Consequently, everyday moral dilemmas offer a more ecologically valid and versatile approach to studying moral judgment, bridging the gap between experimental moral psychology and real-world ethical decision-making. They also broaden the study of moral decision-making by incorporating a framework that distinguishes between altruistic and egoistic choices, expanding the discourse beyond the traditional utilitarian–deontological distinction.

Altruistic and Egoistic Decisions

Moral decision-making is a complex interplay between self-interest and concern for others, often manifesting through egoistic and altruistic motivations. While some decisions are relatively straightforward, like favoring either personal gain or the well-being of others, many moral dilemmas involve conflicting priorities that complicate ethical judgments. In low-conflict situations, choices can be classified as egoistic when they prioritize self-interest and altruistic when they center on helping others. However, from the perspectives of deontology and utilitarianism, these classifications are not always absolute. Deontological frameworks, for instance, do not universally mandate a duty to help, meaning that not all altruistic actions qualify as moral imperatives. Meanwhile, utilitarian evaluations often weigh consequences, making the distinction between personal and moral obligations less clear-cut. Beyond philosophical theory, the debate over human nature has long intrigued scholars in psychology and neuroscience (Haidt, 2007). Rather than existing as entirely separate tendencies, contemporary research suggests that egoism and altruism operate in tandem, shaped by evolved psychological mechanisms and social contexts (Sheninger & Mitra, 2019). Moral choices are influenced by a range of factors, including emotional intensity, time constraints, and reputational considerations (Volz et al., 2017; Zhan et al., 2019). As a result, moral behavior is neither wholly rational nor universally consistent, but rather an adaptive process that balances self-interest with social responsibility. Understanding this dynamic interplay provides insight into the motivations that drive ethical decision-making in everyday life.

Defining Altruism and Egoism "By altruism I mean a motivational state with the ultimate goal of increasing another's welfare." (Batson, 2011, p. 20). The definition of Batson can be broken down into three parts. First, altruism is described as a motivational state, meaning it is not merely a spontaneous reaction or impulse but a goal-directed psychological force. This motivation involves an internal drive to achieve a specific outcome, which in this case is improving another person's well-being. As a goal-directed state, altruistic motivation persists even in the face of obstacles, prompting individuals to seek alternative ways to reach the intended goal. The second key aspect of altruism is that it is defined by its ultimate goal. An ultimate goal is pursued as an end in itself rather than as a means to another objective. This

distinguishes altruistic motivation from instrumental goals, where an action is taken to serve a broader or indirect purpose. In the case of altruism, the improvement of another's welfare is not undertaken for personal gain or external rewards but is the primary and non-negotiable objective. Finally, altruism specifically aims at increasing another's welfare. This involves imagining a desirable change in another person's situation and experiencing an internal force to bring about that change. Altruistic motivation differs from unintended consequences or actions that may incidentally benefit others. Rather, it requires an active commitment to improving another's well-being. The focus on another's welfare as the defining purpose of the motivation is what separates altruism from other forms of prosocial behavior that may be driven by self-interest or external pressures.

Egoism, on the other hand, describes a "motivational state with the ultimate goal of increasing one's own welfare"(Batson, 2011, p. 20). Like altruism, egoism is goal-directed and involves an active psychological force, but instead of being driven by concern for another's well-being, it is fundamentally oriented toward self-benefit, whether through personal gain, social recognition, avoidance of discomfort, or other forms of self-serving motivation (Batson, 2011).

Interaction Between Egoism and Altruism This dichotomy has been widely examined across psychology, neuroscience, and evolutionary biology to uncover the mechanisms underlying moral decision-making. While traditionally viewed as opposing forces, recent research suggests that altruism and egoism function as context-dependent motivations, shaped by self-relevance, reputational concerns, and moral trade-offs. Moral choices often reflect egoistically biased altruism, where individuals are willing to forgo personal rewards to spare others from harm but are far less inclined to endure harm themselves for the benefit of others. While altruistic tendencies exist, they are often moderated by self-interest and risk considerations (Volz et al., 2017). Self-relevance plays a key role in shaping altruistic behavior. Individuals are more likely to act altruistically toward friends and family members than toward acquaintances or strangers. Decision times are shorter when choosing to help close others, suggesting lower cognitive conflict when making prosocial choices for those with strong personal connections. This highlights the role of emotional attachment and social proximity in moral reasoning (Zhan et al., 2019). Reputational concern also influences altruistic behavior. When moral decisions are being observed, individuals behave more altruistically, likely to maintain a positive moral

self-image. This effect is stronger when making decisions about distant others, suggesting that public accountability enhances moral behavior, especially in situations where self-interest might otherwise take priority (Zhan et al., 2019). The type of moral decision influences how people weigh harm and benefit. Individuals are more averse to actively causing harm than to failing to provide a benefit, supporting the idea that moral cognition is sensitive to the nature of the cost involved. People do not apply ethical principles uniformly across different moral dilemmas but adjust their decisions based on the trade-offs between personal risk and prosocial outcomes (Volz et al., 2017). Evolutionary perspectives further suggest that altruism evolved as a cooperative strategy, where helping behaviors ultimately contribute to the survival of social groups (Dawkins, 1976). However, egoistic tendencies also play an adaptive role, as prioritizing one's own well-being can be beneficial in competitive environments (Ayala, 2010). The interaction between these forces suggests that moral decision-making is not solely altruistic or egoistic but rather influenced by a balance of these motivations depending on the situational context.

Moral Foundations Theory as a Foundation of Altruism and Egoism in Moral Decision-Making Congruent with modern research, rather than presenting altruism and egoism as opposing forces, Moral Foundations Theory (MFT) highlights that moral behavior is fluid and context dependent. MFT provides a framework for understanding these decisions by categorizing them into specific moral domains, including care/harm, fairness/unfairness, loyalty/disloyalty, and honesty/dishonesty (Graham et al., 2013) (as explained in further detail in chapter 2.1.4). These foundations shape moral cognition, influencing whether individuals act out of self-interest or prioritize societal well-being.

From a psychological perspective, altruism is typically associated with empathy and compassion, whereas egoism is linked to self-interest and personal gain (Batson et al., 1999). MFT contributes to this understanding by emphasizing moral values that underlie prosocial behavior. Specifically, the care and fairness foundations promote selflessness and justice, fostering altruistic behavior that extends beyond personal relationships to include broader societal concerns (Singer et al., 2019). Conversely, the loyalty, authority, and purity foundations emphasize group cohesion, hierarchy, and moral exclusivity, which often align with egoistic motivations.

Altruistic Behavior and MFT Altruistic behavior is primarily driven by the harm/care and fairness/reciprocity foundations. The harm/care foundation cultivates compassion, motivating individuals to prevent harm and assist others. The fairness/reciprocity foundation supports justice and cooperation, encouraging reciprocal altruism (individuals engage in prosocial behavior based on mutual benefit and fairness) (Graham et al., 2013). These foundations contribute to universal altruism, where moral concern extends to strangers and abstract ethical causes (Singer et al., 2019).

Egoistic Behavior and MFT Egoistic behavior, on the other hand, is influenced by the loyalty/betrayal, authority/subversion, and purity/sanctity foundations. The loyalty foundation fosters selective altruism, wherein individuals prioritize their in-group over outsiders. The authority foundation reinforces obedience to hierarchical structures, sometimes leading individuals to prioritize self-preservation over altruistic actions, particularly when moral responsibility is deferred to an authority figure. The purity foundation is associated with moral exclusivity, where prosocial behavior is conditional on adherence to religious, cultural, or ideological norms (Graham et al., 2013).

Neuronal Measures of Moral Decision-Making

Moral decision-making engages distinct neural networks and has been extensively studied using moral dilemmas, which link different moral choices to specific brain regions. Integrating moral dilemmas into cognitive neuroscience allows researchers to examine how brain activity differs during moral reasoning and emotional conflict resolution. Functional MRI (fMRI) studies typically investigate moral cognition in two ways: (a) individuals making moral choices and (b) judging the moral actions of others (Garrigan et al., 2016). Subsequently, we examine the neural basis of utilitarian, deontological, altruistic, and egoistic decisions. By integrating neuroimaging findings, this chapter explores the interaction between cognitive control and emotional processing in moral choices and how external influences, such as AI-generated moral advice, shape decision-making.

Neural Basis of Utilitarian and Deontological Judgments Moral decision-making is a complex cognitive and affective process involving distinct neural networks. The present

synthesis examines the neurobiological substrates underlying moral reasoning, focusing on differences in brain activation patterns between utilitarian, deontological, altruistic, and egoistic decisions (e.g., Fabre et al., 2024; J. D. Greene et al., 2001). Utilitarian moral judgments, which prioritize aggregate welfare and consequences, engage higher-order cognitive control regions. Functional magnetic resonance imaging (fMRI) studies indicate that utilitarian reasoning correlates with increased activation in the dorsolateral prefrontal cortex (dlPFC) and anterior cingulate cortex (ACC), regions associated with abstract reasoning, cognitive control, and conflict resolution (J. D. Greene et al., 2004). The recruitment of these areas suggests that utilitarian decisions require overriding immediate emotional responses in favor of calculated outcomes. Conversely, deontological judgments, which adhere to moral rules and duty-based reasoning, rely more heavily on emotional processing regions, particularly the ventromedial prefrontal cortex (vmPFC) and amygdala (J. D. Greene et al., 2001; Koenigs et al., 2007). Damage to the vmPFC has been shown to increase utilitarian responses, suggesting its critical role in generating emotional aversion to personal harm and moral violations. Additionally, recent research indicates that deontological reasoning is susceptible to external influence, such as AI-generated moral advice, which can modulate neural activation patterns and response tendencies. AI-generated advice has been shown to affect response times and shift the degree of utilitarian reasoning, likely by influencing cognitive effort allocation and the balance between rational deliberation and emotional responses (Fabre et al., 2024).

Neural Basis of Altruistic and Egoistic Decision-Making Neuroscientific research has identified distinct neural circuits underlying altruistic and egoistic decision-making. Altruistic behaviors, characterized by selfless concern for others, are associated with activation in the ventral striatum, which is linked to reward processing (Hu et al., 2015). This suggests that altruism may be intrinsically rewarding, reinforcing prosocial behavior through dopaminergic mechanisms. Additionally, higher empathic concern correlates with increased activity in the frontoparietal network, particularly the temporoparietal junction (TPJ) and medial prefrontal cortex (mPFC), which are implicated in cognitive perspective-taking and emotional engagement during prosocial decision-making (Eres et al., 2018). The activation of these regions suggests that altruistic choices emerge from an interplay of reward sensitivity, empathy, and cognitive control, reinforcing their prosocial nature. Findings by Rilling and Sanfey

(2011) further support this notion, indicating that altruistic behavior, particularly in social decision-making contexts such as trust and cooperation, is associated with activity in the ventromedial prefrontal cortex (VMPFC) and anterior cingulate cortex (ACC). These regions are linked to social valuation and norm adherence, suggesting that altruistic behavior is reinforced not only by intrinsic reward but also by an individual's perception of fairness and social norms. Moreover, their study highlights that interactions involving reciprocity and fairness activate the anterior insula, reflecting an emotional response to norm violations, further shaping moral decision-making. Conversely, egoistic decision-making engages areas involved in self-interest and goal-directed behavior. The lateral prefrontal cortex (lPFC) and orbitofrontal cortex (OFC) show heightened activity during egoistic choices, reflecting their role in cost-benefit analysis and self-referential processing (Bhuvana et al., 2021). Interestingly, while egoistic decisions also activate reward-related structures similar to altruistic choices, their motivation appears to be externally driven rather than intrinsically rewarding. This suggests that egoistic behavior is more susceptible to extrinsic incentives, such as monetary gains or social reinforcement, rather than the internal satisfaction associated with prosocial actions (Garrigan et al., 2016). According to Rilling and Sanfey (2011), self-interested choices are driven by greater engagement of the lateral prefrontal cortex and amygdala, which suggests that egoistic decisions involve strategic thinking and sensitivity to potential risks or losses. Their research further reveals that egoistic behavior can be reinforced by competitive social settings, where individuals prioritize personal gain over fairness, activating regions linked to reward prediction and aversion to losses.

2.1.6 Advice in Moral Decision Making

Advice as an Additional Source of Information in Moral Decision-Making

Previous chapters have explored different measures and dimensions of morality, examining how individuals assess and apply moral principles in decision-making. Building upon this foundation, the present section introduces an additional variable to the domain of moral measurement: the role of advice in moral decision-making. Specifically, we investigate how the receiving of advice influences ethical judgments and the extent to which different theoretical frameworks can account for these effects. Moral decisions are often complex, requiring

people to navigate conflicting values, ethical norms, and situational pressures. In many cases, individuals do not make these decisions in isolation, but seek, receive, or are exposed to external guidance from advisors, experts, or social influences. The acceptance or rejection of such advice can significantly shape the outcome of moral reasoning, yet the cognitive mechanisms underlying this process remain an important subject of inquiry. We argue that both well-considered and intuitive moral decisions are susceptible to external advice. We build this argument upon several theoretical foundations including dual theory models, the Elaboration Likelihood Model and heuristic decision-making.

The Elaboration Likelihood Model

The Elaboration Likelihood Model (ELM) is a theory by Petty and Cacioppo (1986) that explains how people process persuasive messages and how different factors influence attitude change. According to the model, persuasion occurs through two distinct routes: the central route and the peripheral route. The selection of a particular route depends on the recipient's motivation and ability to engage in effortful cognitive processing. The central route is utilized when an individual possesses both high motivation and high ability to process the message. Motivation is typically driven by personal relevance, issue involvement, or intrinsic interest in the topic, whereas ability is determined by cognitive capacity, prior knowledge, and the availability of time to engage in systematic thinking. When processing occurs via the central route, individuals critically evaluate the quality and strength of the arguments presented. If the arguments are compelling, persuasion leads to stable, enduring attitude change that is resistant to counterarguments. On the other hand, the peripheral route is activated when motivation or ability is low. In such cases, individuals rely on heuristic cues rather than engaging in deep cognitive evaluation of the message's content. Persuasion via this route is less durable and more susceptible to future changes, as attitudes formed through peripheral processing lack the robustness of those shaped through central processing. The distinction between these two routes has significant implications for persuasion and decision-making. For instance, individuals making high-stakes decisions, such as selecting a medical treatment or evaluating scientific evidence, are more likely to engage in central processing, leading to informed and stable attitudes. In contrast, individuals making low-stakes or habitual decisions,

such as routine consumer choices, are more likely to engage in peripheral processing, resulting in more transient and malleable attitudes. Both high-stakes and low-stakes decisions can be transferred into moral dilemmas (see chapter 2.1.5 for further information). Given these two possibilities, either participants are motivated and capable to make moral decisions, or participants are either not capable or not motivated. We argue that both possibilities are prone to advice, but in different ways. The subsequent passages explore both routes and argue how advice influences moral decision-making in distinct manners. The Elaboration Likelihood Model (ELM) serves as a missing link and a logical extension in understanding how external persuasive cues, such as advice, influence moral decision-making within a dual-process framework. Building on the theories of Gigerenzer (2008), Gigerenzer and Gaissmaier (2011), and J. D. Greene et al. (2004, 2008) and others, the ELM explains how advice can shape moral reasoning by engaging both heuristic-based (peripheral route) and deliberative (central route) processing. This integration underscores the role of persuasive input in modulating dual-process moral cognition, demonstrating how external advice interacts with intuitive and analytical moral judgment mechanisms. While moral identity displays the motivational aspect of the ELM, wisdom is likely more closely related to the ability aspect of the central route.

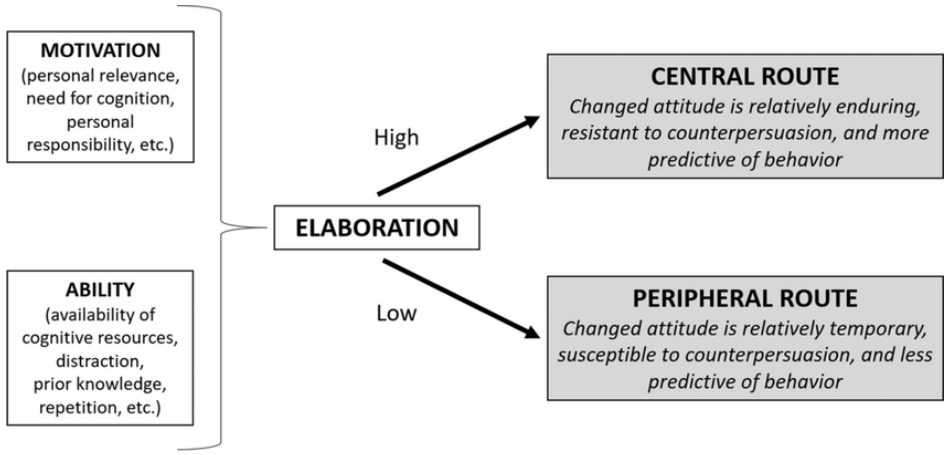


Figure 2.1: Elaboration Likelihood Model (ELM) as presented in Petty and Cacioppo (1986).

Central Route Together with a thoughtful and lasting moral decision, which involves weighing different alternatives and integrating persuasive advice, we introduce two key concepts that we expect to increase the likelihood of deliberate moral reasoning. We argue that Moral Identity

Theory (Aquino & Reed, 2002) and Wisdom (Staudinger & Glück, 2011) are particularly relevant to thoughtful central route processing within the Elaboration Likelihood Model (ELM) in the context of advice-dependent moral decision-making.

Moral Identity Theory Moral Identity Theory, introduced by Aquino and Reed (2002), describes how morality and the aspiration to be a moral person function as a core aspect of social identity. Just as people define themselves through various social categories (e.g., professional identity, religious identity), a person's moral identity may be associated with certain beliefs, attitudes, and behaviors, particularly when that identity is highly self-important. According to Aquino and Reed, moral identity functions as a self-concept structure, where moral traits (e.g., fairness, kindness, honesty) become central to an individual's sense of self. Individuals with a strong moral identity experience moral concerns as personally significant, leading them to engage in deliberate moral reasoning rather than relying on external influences or heuristic processing.

Moral Identity Theory as Motivation in Central Route Processing This intrinsic commitment to morality increases the likelihood of central route processing (Petty & Cacioppo, 1986), as these individuals are intrinsically motivated to analyze moral dilemmas critically, assessing arguments based on logic, coherence, and ethical soundness rather than superficial cues. Empirical research supports the idea that moral identity strength is associated with greater moral engagement and ethical decision-making (Aquino et al., 2009). The self-imposed expectation to be a moral person effectively compels individuals to engage deeply with the moral advice they receive, carefully considering its arguments and integrating them into their own moral reasoning process.

Advice-Taking In the context of advice-taking, individuals with a strong moral identity are more likely to scrutinize the content and ethical validity of moral advice, rather than being swayed by the advisor's social status or authority. From the perspective of Moral Identity Theory, those who strongly identify as moral persons are intrinsically motivated to process moral advice through the central route of persuasion. Because their moral self-concept is highly salient, they are less likely to dismiss or superficially accept advice based on heuristics.

Instead, they engage in thoughtful elaboration, critically assessing the ethical validity, logical coherence, and long-term implications of the arguments presented. Thus, the desire to uphold one's moral identity acts as a cognitive and motivational driver, increasing the likelihood of central processing in moral decision-making. However, situational cues have the power to either reinforce or deactivate the moral identity within a working self-concept (Aquino et al., 2009). When applied to persuasive advice as an external situational cue, we argue that individuals are susceptible to such external influences, particularly when following the central route of the ELM. That is, even though moral identity strengthens deep reasoning, the framing and presentation of persuasive advice can still shape the way individuals engage with moral arguments.

Wisdom In addition to moral identity, wisdom plays a critical role in thoughtful moral decision-making, enabling individuals to analyze moral dilemmas with depth, recognize complexity, and engage in deliberative ethical reasoning (Baltes & Staudinger, 2000). While moral identity is closer to the motivational aspect of central route processing in moral persuasion, wisdom is more closely related to the ability component, as it equips individuals with the cognitive and reflective skills necessary for systematic moral evaluation. Wisdom is commonly understood as a state of mind and behavior that integrates intellectual, affective, and motivational dimensions of human functioning. It encompasses reflective thinking, intellectual humility, and integrative reasoning, all of which enhance an individual's ability to process moral information through the central route, rather than relying on superficial cues.

Wise individuals tend to engage in critical analysis of moral arguments, evaluating their logical soundness rather than accepting them based on authority or emotional appeal. They are also more adept at recognizing complexity and uncertainty in ethical dilemmas, avoiding rigid, black-and-white thinking. Furthermore, wisdom fosters a long-term perspective, encouraging individuals to prioritize ethical consequences over impulsive, emotionally driven decisions. This deep moral reasoning aligns with the seven properties of wisdom outlined by Baltes and Staudinger (2000), including superior knowledge, judgment, and advice, as well as awareness of the limitations of knowledge and the uncertainties of the world.

By combining cognitive complexity, self-reflection, and emotional regulation, wise individuals are equipped to carefully evaluate moral dilemmas rather than defaulting to heuristics or

external influences (Ardelt, 2004).

Wisdom as Moral Ability in Central Route Processing Within the Elaboration Likelihood Model (ELM), wisdom plays a crucial role in determining an individual's ability to process moral advice systematically. While moral identity provides the motivational drive to engage with ethical dilemmas, wisdom ensures that individuals have the cognitive and reflective resources necessary to analyze moral information deeply. Wise individuals excel at distinguishing substantive moral arguments from persuasive rhetoric, making them more resistant to heuristic-based reasoning. Their decision-making is characterized by a balance between rational analysis and ethical intuition, ensuring that choices are not only logically coherent but also morally grounded. A key feature of wisdom is metacognitive reflection, which enables individuals to question their own biases, reassess assumptions, and critically evaluate competing perspectives before forming ethical judgments. Empirical research supports the idea that wisdom enhances ethical decision-making and fosters thoughtful elaboration in moral reasoning (Grossmann et al., 2010; Staudinger & Glück, 2011). When confronted with persuasive moral advice, wise individuals are more likely to systematically evaluate its validity, integrate multiple viewpoints, and arrive at well-reasoned moral stances that are less susceptible to external influence.

Wisdom in Moral Advice-Taking Wisdom shapes how individuals interpret, evaluate, and integrate external moral advice. Unlike those who may be influenced by social consensus, authority figures, or emotional appeals, wise individuals approach moral guidance with a balance of skepticism and openness. They carefully assess the logical strength, ethical soundness, and practical consequences of the advice before deciding whether to accept or reject it. From an ELM perspective, wisdom enhances central route processing in two primary ways. First, it reduces reliance on heuristics, encouraging individuals to base their moral decisions on reasoned analysis rather than external credibility cues. Second, it strengthens metacognitive awareness, allowing individuals to reflect on their own decision-making process, question their cognitive biases, and ensure their judgments are well-justified.

Conclusion However, while central route processing provides a framework for deliberate and reflective moral reasoning, it is not always the dominant mode of decision-making. There are numerous situational and cognitive constraints that may prevent individuals from engaging in effortful moral deliberation. In many cases, motivation or cognitive ability, which are both essential for central processing, are not fully present. Time pressure, cognitive overload, distraction, emotional distress, or a lack of expertise can significantly reduce an individual's capacity or willingness to systematically evaluate moral arguments in their entirety. When either motivation or ability is insufficient, individuals are more likely to rely on heuristic shortcuts rather than engaging in deep moral reflection. Congruently, Gigerenzer (2010) argues that most moral behavior is based on heuristics rather than on extensive reasoning or systematic ethical evaluation. This is likely due to real-life circumstances and numerous distractions. Instead of engaging in deliberate moral analysis, individuals often depend on cognitive shortcuts, social cues, and intuitive moral instincts to guide their ethical decisions. Heuristics allow for fast, efficient decision-making, particularly in environments where individuals face high uncertainty, time constraints, or limited cognitive resources. Given these considerations, we acknowledge that while central route processing plays a crucial role in thoughtful moral decision-making, heuristic-based decision-making is the more frequent mode of moral reasoning in everyday life.

Peripheral Route In the following passage, we place a special emphasis on the peripheral route of the Elaboration Likelihood Model (ELM), which focuses on heuristic processes and the influence of external cues on decision-making. This section explores how persuasive advice shapes moral judgments through peripheral processing, where decisions are guided by heuristics, social influence, and emotional appeals rather than systematic moral evaluation. Research supports the idea that cognitive load increases reliance on heuristics. J. D. Greene et al. (2008) found that when individuals experience cognitive overload, their ability to engage in deep moral reasoning declines, making them more likely to depend on heuristic-based decision-making. In such cases, advice-seeking acts as a cognitive bypass, helping individuals reach moral judgments more efficiently without engaging in extensive deliberation.

The Elaboration Likelihood Model highlights how individuals process persuasive information

through either a cognitively demanding central route or a more effortless peripheral route. The likelihood of elaboration depends on motivation and ability, meaning that people often rely on simple cues like source credibility, consensus, or emotional resonance, when cognitive resources are limited or motivation is low. These peripheral cues operate as heuristic shortcuts that guide judgments with minimal deliberation, mirroring the principles of Type 1 processing outlined in chapter 2.1.4. To further understand the cognitive mechanisms underlying such low-effort processing, the following section turns to the concept of heuristics and explores how they function as adaptive effort-reduction strategies in moral decision-making, including how advice itself can serve as a heuristic cue.

Heuristics in Moral Decision-Making

The following passage is twofold. First, we aim to introduce the concept of heuristics and heuristic decision-making, integrating different heuristics into a broader concept of effort reduction principles. Second, we elaborate on which heuristics fit the interpretational space of advice and why advice ultimately functions as a principle of effort reduction.

Definition and Origins of Heuristics Heuristics are mental shortcuts or "rules of thumb" that simplify decision-making under uncertainty by focusing attention on the most salient or predictive cues while ignoring others (Gigerenzer & Gaissmaier, 2011). The heuristics-and-biases program first demonstrated that, under conditions of uncertainty, people rely on efficient shortcuts (e.g. availability and representativeness) that often produce systematic judgment errors (Tversky & Kahneman, 1973, 1974). A unifying account of these effects is attribute substitution: when a target attribute is difficult to evaluate, individuals unconsciously replace it with a more accessible heuristic attribute (Frederick, 2005; Kahneman & Frederick, 2002). In moral contexts, such substitutions may involve intuitive cues like authority, consensus, or sanctity, which act as simplified proxies for complex moral reasoning (Sunstein, 2005). These mechanisms allow for rapid judgments but can also introduce predictable biases.

From the perspective of bounded rationality, heuristics emerge as adaptive solutions to the limitations of human cognition. As H. A. Simon (1990, p.7) famously stated, "human rational behavior is shaped by a scissors whose two blades are the structure of task environments

and the computational capabilities of the actor“. Because time, information, and cognitive resources are finite, individuals rely on approximate strategies to make decisions that are good enough given their constraints. Building on this view, the ecological rationality approach reframed heuristics not as sources of error but as context-sensitive strategies that can be highly effective when they exploit environmental regularities (Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 2001). In certain environments, heuristics may even outperform more complex algorithms by emphasizing the most valid cues and disregarding irrelevant noise.

As the demands on limited cognitive resources increase, people are more likely to adopt methods that minimize computational effort. Heuristics can therefore be understood as strategies that implement principles of effort reduction and simplification (Shah & Oppenheimer, 2008). They enable decision-makers to process information efficiently while balancing the trade-off between accuracy and effort. Importantly, heuristics are neither inherently flawed nor universally optimal; their effectiveness depends on the fit between cognitive mechanisms and environmental structure. Within moral cognition, heuristics form the intuitive substrate of moral judgment, allowing individuals to navigate complex ethical decisions through rapid, affectively charged, and socially informed evaluations.

Examples of Heuristics and Their Relevance for Moral Decision-Making Heuristics manifest in numerous forms, each reflecting a specific strategy for simplifying complex judgments. Among the most studied examples is the less-is-more heuristic, which suggests that in some cases, having less information can lead to better decisions than having more. When additional data introduce noise rather than insight, individuals may perform more accurately by relying on a smaller, more diagnostic subset of cues (Gigerenzer & Gaissmaier, 2011). In moral decision-making, this principle implies that relying on concise advice or clear moral guidance can prevent overthinking or paralysis when faced with competing ethical considerations.

Another prominent mechanism is the availability heuristic, the tendency to judge the frequency, likelihood, or importance of events by how easily relevant examples come to mind (Tversky & Kahneman, 1973). In moral contexts, this means that vividly imagined or emotionally charged scenarios, for instance graphic stories of individual victims, striking cases of injustice, or salient media reports of wrongdoing, are retrieved from memory more fluently than abstract, statistical harms. The subjective ease of recall is then misused as

an informational cue: what comes to mind quickly is experienced as more probable, more widespread, and morally more urgent than less accessible harms. As a result, singular, concrete cases can weigh more heavily in moral judgment than objectively larger but less imaginable harms, and recent or emotionally intense events can disproportionately shape which issues feel morally pressing, even when they are not representative of the overall risk landscape.

The one-reason decision-making strategy further exemplifies how individuals economize cognitive effort by focusing on a single decisive cue while ignoring all others (Gigerenzer & Gaissmaier, 2011). In moral contexts, such a cue might involve prioritizing harm avoidance or fairness, depending on the individual's moral orientation. Advice often leverages this mechanism e.g. by highlighting one compelling argument, such as "the ends justify the means" or "it is wrong to lie", thereby enabling the decision-maker to resolve complex moral dilemmas by focusing on a single evaluative dimension rather than engaging in multi-attribute deliberation.

Closely related are fast and frugal heuristics, minimal-information strategies that achieve efficient and often surprisingly accurate judgments by aligning with environmental regularities (Gigerenzer, 2008). In moral reasoning, such heuristics may take the form of simple decision rules (e.g., "do no harm" or "help those in need") that are culturally transmitted and rapidly applied. These heuristics demonstrate how moral behavior can be adaptive without requiring extensive reasoning. Individuals can act morally appropriate in most circumstances by following intuitive, socially reinforced guidelines.

The anchoring heuristic illustrates how initial information can disproportionately shape subsequent judgments. When people rely too heavily on a given reference point, later adjustments tend to be insufficient (Tversky & Kahneman, 1974). In moral decision-making, advice can act as such an anchor: once an advisor provides a recommendation, individuals often evaluate subsequent arguments in relation to that initial position, resulting in a bias toward the advised judgment even when counterevidence is presented.

These heuristics illustrate the adaptive yet fallible nature of human moral cognition. They enable individuals to reach judgments quickly and with limited effort, but they also render (moral) reasoning susceptible to contextual framing and social influence. This dual nature is particularly evident when considering the role of advice in moral contexts.

Advice as a Heuristic in Moral Decision-Making Navigating the complexity of moral dilemmas often exceeds the cognitive capacity or motivational resources of the individual. In such situations, seeking advice can function as an adaptive heuristic that reduces cognitive load, clarifies ambiguity, and provides a socially validated course of action. Rather than engaging in laborious moral reasoning, individuals frequently defer to the judgment of trusted advisors, allowing them to substitute another's reasoning for their own. This process mirrors the general logic of heuristics: it simplifies decision-making by selectively reducing effort while maintaining sufficient accuracy for practical purposes.

Advice-seeking can thus be understood as a compound heuristic that integrates several of the mechanisms described above. It aligns with the less-is-more heuristic by filtering out excess information and focusing on concise moral cues, and with the availability heuristic by rendering certain moral aspects more salient through the advisor's framing. Similarly, it parallels "one-reason decision-making" when the advisor highlights a single decisive argument, thereby channeling attention toward one dominant evaluative dimension. Advice also exhibits the characteristics of fast and frugal heuristics, as it enables individuals to make quick, confident moral judgments by relying on socially endorsed rules of thumb. Also, advice acts as an anchor that shapes subsequent evaluations, creating a reference point from which individuals insufficiently adjust even when alternative information becomes available. These examples are only a brief selection of heuristics that can be used to explain the use of advice in decision-making.

In this sense, advice-seeking embodies the core principles of heuristic processing. It serves as a cognitive and social shortcut that reduces informational complexity, preserves mental resources, and allows for efficient moral judgment. At the same time, its reliance on social and contextual cues makes it vulnerable to distortion: advice can amplify certain moral perspectives while suppressing others, leading to conformity, persuasion, or bias. Conceptualizing advice as a heuristic offers a framework for examining how individuals navigate moral uncertainty by relying on socially embedded shortcuts rather than extended deliberation.

Advice as a Social Heuristic in (Moral) Decision-Making Advice-seeking can also be understood as a social heuristic, where individuals defer to others' judgments as a trust-based shortcut rather than engaging in complex ethical deliberation (Gigerenzer & Gaissmaier, 2011).

Rather than relying solely on personal reasoning, people often look to external sources for moral guidance, simplifying their decision-making process. Moral decisions frequently depend on social heuristics, as individuals tend to conform to group norms rather than applying strict ethical principles in isolation (Gigerenzer, 2010). In moral decision-making, people are more likely to trust well-known experts or institutions over unfamiliar sources, simplifying their decision process (del Campo et al., 2016). Social heuristics reduce cognitive effort while promoting socially accepted choices, reinforcing moral conformity and guiding individuals toward decisions that align with collective values. Research suggests that individuals rely on social heuristics to imitate group behavior, prioritizing group cohesion over independent ethical deliberation (Gigerenzer, 2010). This tendency is particularly strong in moral dilemmas involving sacred values, where decision-making becomes more emotionally and cognitively demanding (Hanselmann & Tanner, 2008). When a sacred value (e.g., human life) conflicts with a secular value (e.g., economic gain), advice from trusted sources can reinforce an immediate moral reaction, leading to faster and more confident decisions. In cases where two sacred values are in conflict, such as choosing between saving one life over another, advice-seeking helps structure moral reasoning, making it easier to resolve complex dilemmas. Beyond individual moral choices, advice-seeking fosters cooperation, as both advice-seeking and cooperation function as social heuristics, allowing individuals to align with group norms and make efficient decisions that benefit collective well-being. Studies have shown that time pressure increases cooperation, suggesting that prosocial behavior is often guided by intuitive social heuristics rather than extensive deliberation (Rand et al., 2014). Similarly, Todd et al. (2008) found that cooperation is reinforced by social heuristics, supporting the idea that advice-seeking serves as a mechanism for moral alignment within a group. When individuals seek advice, they often receive confirmation that cooperative behavior is the expected or beneficial choice, reinforcing moral conformity and ensuring that decisions align with shared ethical standards.

We have now considered multiple heuristics that shape the influence of persuasive advice on moral decision-making. In practice, these heuristics often overlap, as there are theoretically as many heuristics as there are individual cases, or as As Shah and Oppenheimer (2008, p. 215) stated that there are "*as many heuristics as there are cues for judgment in the world*".

Consequently, we introduce the concept of effort reduction as a unifying heuristic principle, arguing that advice satisfies multiple effort-reduction strategies simultaneously.

Effort Reduction as a Holistic Singular Heuristic Understanding heuristic decision-making requires recognizing that no single heuristic accounts for all cognitive shortcuts used in different contexts. Instead, heuristics operate as adaptive mechanisms that streamline decision-making by selectively reducing cognitive effort. The effort reduction framework by Shah and Oppenheimer (2008) provides a structured approach to analyzing how heuristics achieve this by targeting specific cognitive processes. It identifies five fundamental methods of effort reduction, each addressing a particular way in which cognitive load can be minimized, thereby enhancing efficiency in judgment and decision-making.

(1) Examine fewer cues. This strategy reduces cognitive effort by limiting the number of factors considered when making a decision. Instead of analyzing all available information, these heuristics focus on the most critical or predictive cue while disregarding others. This selective attention simplifies the decision process by reducing the cognitive load associated with comparing multiple attributes simultaneously. A variety of heuristics employ this strategy, either by prioritizing a single most valid cue or by systematically eliminating less relevant ones. While some heuristics require sequential cue evaluation, others, such as lexicographic models, enforce strict prioritization rules that prevent unnecessary information processing. The effectiveness of this method depends on the ecological validity of the selected cues, as disregarding relevant information may lead to suboptimal choices.

(2) Reduce retrieval and storage difficulty. Decision-making often necessitates recalling, organizing, and comparing multiple pieces of information, which can impose significant cognitive demands. To mitigate this effort, heuristics employ strategies that either simplify memory retrieval or reduce the complexity of stored representations. This can be achieved through categorical encoding, where detailed quantitative differences are replaced with broad classifications, or by relying on previously stored judgments rather than recalculating cue values. Some heuristics further reduce memory load by operating on relative comparisons (e.g., "better than" rather than exact numerical differences), which minimizes the need for detailed

mental representations. The efficiency of this approach is contingent upon the accessibility and reliability of the retrieved information, as reliance on easily accessible but biased memory structures can introduce systematic errors in judgment.

(3) Simplify weighting principles. This approach reduces cognitive effort by eliminating or standardizing the weighting of decision-relevant factors. In a normative decision model, cues are typically assigned differential weights based on their predictive validity, requiring additional cognitive effort to establish and apply these weights. To avoid this complexity, some heuristics assign equal weight to all cues, thereby simplifying the integration process and reducing computational demands. Other strategies base cue weighting on prior success or frequency of use, implicitly reinforcing previously effective decision rules. Heuristics that employ this method tend to function well in environments where the relative importance of cues remains stable, but they may be less effective in contexts requiring dynamic adjustments of weighting principles.

(4) Integrate less information. This strategy restricts the amount of data that must be processed when making a decision. Instead of aggregating multiple attributes into a comprehensive evaluation, heuristics using this method apply threshold-based decision rules, terminating the decision process once a satisfactory option is identified. This approach eliminates the need for exhaustive comparisons and allows decision-makers to focus only on whether an alternative meets predefined criteria. Some heuristics further reduce information integration by sequentially eliminating alternatives that fail to satisfy minimal requirements, thereby limiting the scope of the final evaluation. While this method improves efficiency, its effectiveness depends on the appropriateness of the threshold applied, as overly strict or lenient criteria may lead to premature or suboptimal choices.

(5) Examine fewer alternatives. Decision-making is inherently more complex when multiple alternatives must be compared simultaneously, as each additional option introduces further cognitive load. Heuristics employing this method reduce effort by restricting the decision space, either by sequentially eliminating alternatives that do not meet critical criteria or by employing pairwise comparisons where only two options are evaluated at a time. Some

heuristics further refine this process by structuring decision tasks to promote early rejection of inferior alternatives, ensuring that only a limited subset is considered in later stages. The efficiency of this method is contingent upon the decision-maker's ability to correctly identify and discard less viable alternatives early in the process.

Why Advice Fits as an Effort Reduction Principle We argue that advice-seeking can be understood as a heuristic that functions within the framework of effort reduction, as proposed by Shah and Oppenheimer (2008). Rather than engaging in extensive cognitive deliberation, individuals rely on advice as a mental shortcut to simplify complex decision-making. Evaluating advice utilization through the lens of effort reduction reveals that it aligns with several core mechanisms designed to minimize cognitive effort while maintaining decision efficiency.

First, advice facilitates examining fewer cues, as it enables decision-makers to focus on a single decisive factor while disregarding all competing information. Instead of evaluating multiple influences, individuals concentrate solely on the most salient cue, preventing decision paralysis and reducing cognitive effort.

Advice reduces the difficulty of retrieving and storing cue values, sparing individuals the need to recall, compare, or analyze multiple pieces of information. By presenting a direct recommendation, advice removes the burden of memory-based retrieval and simplifies the choice to a readily available option.

Advice-taking contributes to simplifying the weighting principles for cues by eliminating the need for relative trade-offs. Binary advice, in particular, assigns absolute importance to a single factor, removing the effort required to balance multiple criteria or assess competing influences. This makes the decision feel automatic, as it is framed as self-evident rather than requiring complex deliberation.

Advice facilitates integrating less information, streamlining the judgment process into a singular directive. The decision-maker does not need to synthesize different perspectives or weigh conflicting evidence, thereby reducing cognitive processing demands.

Lastly, advice-seeking aligns with examining fewer alternatives by collapsing the decision space into a simple "yes or no" framework. Presenting a single option as the most viable choice discourages further deliberation and reinforces a sense of certainty while limiting hesitation

and doubt. This extreme reduction of decision complexity maximizes cognitive efficiency but also introduces the risk of oversimplification. By narrowing choices to a singular perspective, important contextual nuances may be overlooked, potentially compromising adaptability in complex decision environments.

Conclusion The preceding sections have argued that many heuristics can be understood as implementing a small set of effort-reduction strategies, and that the framework by Shah and Oppenheimer (2008) provides a useful lens for organizing these strategies. Rather than treating each heuristic as a separate phenomenon, the effort-reduction perspective highlights their common function: to make satisfactory decisions possible under constraints of limited time, information, and cognitive resources. Within this framework, advice-seeking emerges not as a marginal add-on to individual judgment, but as a particularly powerful instantiation of effort reduction in (moral) decision-making.

Conceptualizing advice as an effort-reduction principle clarifies why it is so attractive in complex moral contexts. Accepting advice allows decision-makers to consider fewer cues, retrieve and integrate less information, dispense with elaborate weighting schemes, and narrow the set of perceived alternatives to a simple, binary choice. In doing so, advice operates as a composite heuristic that bundles several effort-reduction mechanisms into a single social cue. It compresses informational complexity into an actionable recommendation, thereby lowering the motivational and cognitive thresholds for reaching a moral judgment.

At the same time, this analysis also makes visible the potential costs of such efficiency. Precisely because advice streamlines moral cognition so effectively, it can encourage premature closure, anchor judgments on an externally provided option, and obscure relevant nuances of the situation. When advice is accepted uncritically, the underlying effort-reduction strategies may facilitate conformity or bias rather than reflection and autonomy. Understanding advice as a composite effort-reduction principle therefore underscores both its adaptive role in managing moral complexity and its capacity to systematically shape, and sometimes distort, moral decision-making.

2.1.7 Summary

This chapter provided a systematic and interdisciplinary overview of morality, integrating philosophical, psychological, neuroscientific, and decision-theoretical perspectives to establish a comprehensive foundation for the empirical investigation of moral decision-making. Beginning with classical philosophy, it traced the evolution of ethical thought from virtue-based approaches, through deontological and consequentialist traditions, to modern psychological theories that emphasize intuition, emotion, and social context in moral cognition.

Early philosophers such as Aristoteles et al. (2011), Kant (1998), and Mill (1966) established enduring frameworks for understanding morality in terms of character, duty, and outcomes. These perspectives remain foundational in contemporary discussions, particularly in the analysis of deontological and utilitarian judgments in moral dilemmas. Building upon these philosophical roots, psychological theories of moral development, including those of Piaget (1934) and Kohlberg (1985), proposed stage-based models that highlighted the role of cognitive maturation in moral reasoning. However, these rationalist and linear conceptions were later challenged for their cultural bias, lack of emotional depth, and limited ecological validity.

Modern approaches to morality, especially those proposed by Graham et al. (2011, 2013, 2016), J. Greene and Haidt (2002), J. D. Greene et al. (2001, 2004, 2008, 2009), J. D. Greene (2014), Haidt (2001, 2007, 2008, 2013), Haidt and Joseph (2004), and Haidt and Kesebir (2010) shifted the focus toward intuitive and affective processes. The Social Intuitionist Model (SIM) emphasized that moral judgments typically arise from rapid, emotion-based intuitions, with reasoning primarily serving a post-hoc justificatory role. Moral Foundations Theory (MFT) further specified the content of these intuitions by identifying distinct, evolutionarily grounded domains such as care, fairness, loyalty, authority, sanctity, and liberty. Together with the New Synthesis of Morality, these frameworks situated moral cognition within a broader evolutionary and social context, proposing that morality primarily functions to regulate selfishness, foster cooperation, and bind individuals into cohesive groups.

Dual-process models provided a crucial cognitive architecture for these theories by distinguishing between fast, automatic, heuristic-driven processes and slower, deliberative, effortful reasoning. In moral contexts, intuitive deontological responses are generally associated with emotional brain networks, whereas utilitarian reasoning relies on cognitively demanding control

processes. This distinction was supported by extensive neuroscientific evidence linking different forms of moral judgment to specific neural networks involving the amygdala, prefrontal cortex, anterior cingulate cortex, and related networks.

The chapter also explored how morality is operationalized in empirical research. Moral dilemmas, particularly sacrificial cases such as the Trolley problem by Foot (1967), were presented as central measurement tools for studying moral cognition. While these scenarios offer experimental control and theoretical clarity, they also suffer from limitations in ecological validity and psychological realism. As a response, everyday moral dilemmas were introduced as a more contextually grounded and flexible method for examining ethical decision-making in real-world contexts (Singer et al., 2019). These were further complemented by distinctions between altruistic and egoistic choices, which capture the motivational dynamics underlying moral behavior.

A central contribution of this chapter lies in its integration of persuasion theory and heuristic decision-making into the study of moral judgment. The Elaboration Likelihood Model (ELM) was applied to moral contexts to distinguish between central (systematic) and peripheral (heuristic) processing of moral information. Moral identity and wisdom were introduced as individual-level factors that increase motivation and ability for central processing, whereas time pressure, cognitive load, and emotional salience promote reliance on peripheral cues.

Heuristics were then framed as adaptive tools of bounded rationality, enabling individuals to make efficient decisions under uncertainty. Mechanisms such as the availability heuristic, anchoring, one-reason decision-making, and fast-and-frugal strategies were presented as general cognitive shortcuts that also operate in moral contexts. These heuristics reduce cognitive effort by simplifying information processing, narrowing attention, and substituting complex evaluations with intuitive cues.

Advice was conceptualized as a specialized heuristic through which individuals rely on external input in moral decision-making. Advice satisfies key criteria of effort reduction by limiting the number of cues, easing memory demands, simplifying weighting and integration, and restricting the range of considered alternatives. By serving as an anchor, a credibility cue, and a source of social validation, advice influences moral judgment while reducing cognitive demands.

This chapter argues that moral decision-making is shaped by cognitive constraints, emotional intuitions, and social influences. Moral judgments are not exclusively the product of deliberate reasoning, but are embedded within a network of heuristics and interpersonal guidance mechanisms. Advice represents one such effort-reduction mechanism, enabling individuals to navigate complex moral situations efficiently, while also introducing potential sources of bias and susceptibility to influence.

This theoretical foundation establishes a clear and necessary bridge to the core focus of the dissertation: the study of Artificial Moral Advisors (AMAs) and their potential to influence human moral decision-making. If morality is intuitive, socially grounded, and shaped by effort-reducing shortcuts, then artificial systems offering moral advice may meaningfully alter how individuals judge, decide, and attribute moral responsibility. The subsequent chapters first turn to the broader domain of Artificial Intelligence, outlining its conceptual foundations and societal implications, before introducing Artificial Moral Advisors (AMAs) as a specific case. These strands are then integrated in the following chapter to formulate the overarching aim and structure of the present dissertation.

2.2 Artificial Intelligence

Artificial Intelligence (AI) has emerged as one of the most transformative technologies of the twenty-first century, revolutionizing industries and redefining the way humans interact with machines. Its ability to analyze vast amounts of data, recognize patterns, make predictions, and automate complex tasks has led to groundbreaking advancements in fields such as healthcare, finance, manufacturing, transportation, defense, and scientific research (Berman et al., 2023; V. Kumar, 2024; Sengar et al., 2024; Weng et al., 2024). This chapter aims to provide a definition for Artificial Intelligence, an overview of different application fields, as well as special emphasis on the perception of AI in the domain of moral decision-making.

2.2.1 Introduction to Artificial Intelligence

Artificial Intelligence (AI) is a rapidly evolving field that plays a transformative role in modern technology. While there is no universally accepted definition, AI is broadly understood as systems that analyze their environment, make decisions, and take actions (with some degree of autonomy) to achieve specific goals (European Commission, 2019; European Court of Auditors, 2024). These systems can be purely software-based, such as search engines and speech recognition tools, or embedded in hardware, including robotics and autonomous vehicles. A common distinction in AI research is between AI as an imitation of human intelligence and AI as rational decision-making. The latter approach defines AI as machines that " function appropriately and with foresight in their environment" (Nilsson, 2009, p.13), emphasizing its ability to process information and optimize decisions without necessarily replicating human cognition. AI can be built using symbolic rules, numerical models, or machine learning techniques that allow it to adapt its behavior based on data and experience (European Commission, 2019). Beyond technical definitions, AI is also recognized for its societal impact. The European Court of Auditors (2024, p.4) describes Artificial Intelligence (AI) as "a technology that comes with a promise to transform economies, boost growth and address societal challenges, but it also carries inherent safety risks and significant potential for economic and societal disruption". As AI advances, discussions increasingly focus on its ethical implications, including fairness, transparency, and accountability. Consequentially, AI is best defined as a field of computer science focused on developing systems capable of learning,

reasoning, and decision-making to perform tasks traditionally requiring human intelligence. Its definition continues to evolve as research progresses and AI integrates further into society (European Commission, 2019; European Court of Auditors, 2024; Nilsson, 2009; Sheikh et al., 2023).

2.2.2 Definition

The concept of Artificial Intelligence (AI) is inherently complex and resists a single, universally accepted definition, prompting ongoing debates about its scope, function, and implementation. Across disciplines, scholars and policymakers emphasize different aspects (ranging from cognitive emulation and learning capacity to autonomy, rationality, and societal embedding). This definitional diversity is not a weakness but reflects AI's dual nature as both a technological and socio-cognitive phenomenon that challenges conventional understandings of intelligence, agency, and human–machine interaction (Caluori, 2024; Sheikh et al., 2023).

Early conceptualizations of AI were grounded in the ambition to replicate human intelligence. When John McCarthy (1956) introduced the term, he defined AI as "the science and engineering of making intelligent machines," highlighting the goal of designing systems capable of reasoning, problem-solving, and learning (McCarthy, 1956, p.2). Subsequent developments expanded this perspective, replacing symbolic reasoning with adaptive, data-driven models such as machine learning and neural networks. As a result, the term AI now encompasses a broad spectrum of systems that vary in autonomy, complexity, and scope from rule-based expert systems to generative, self-learning architectures (Collins et al., 2021; Ng et al., 2021).

A comprehensive overview of definitional approaches is provided by the WRR report Mission AI: The New System Technology (Sheikh et al., 2023). The authors note that AI definitions range from equating AI with any algorithmic process to more restrictive conceptions of machines that fully replicate human intelligence. At one extreme, broad algorithmic definitions are deemed analytically unhelpful, as they encompass trivial computation such as calculator operations. At the opposite end, definitions tied to Artificial General Intelligence (AGI), systems capable of reproducing all human intellectual abilities, are dismissed as too idealized, since such systems do not yet exist. Between these poles lies a functional definition that has gained broad acceptance: "systems that display intelligent behaviour by analysing their

environment and taking actions—with some degree of autonomy—to achieve specific goals." This formulation, developed by the European Commission (2019, p.1)'s High-Level Expert Group on AI (AI HLEG), distinguishes AI from general digital technologies while remaining flexible enough to accommodate future advancements. It captures two key features that form the foundation of contemporary AI discourse: autonomy and goal directed rationality (Sheikh et al., 2023).

The European Commission (2019) expands this definition to specify how AI systems operate. These systems, designed by humans, are assigned complex goals and function in either digital or physical environments. They perceive their surroundings through data acquisition, reason and process information to derive knowledge, and then decide on and execute the actions most likely to achieve the given objective. AI systems thus combine perception, reasoning, and actuation in a closed adaptive loop, enabling them to learn from feedback and refine their performance. Rationality, rather than intelligence per se, serves as the conceptual foundation of this framework: AI systems are not required to "think" like humans but to select the optimal action within the constraints of available information and resources (Russell & Norvig, 2022). The AI HLEG further differentiates between several core scientific domains that underpin these capabilities. It highlights machine learning, which enables systems to adjust their behaviour on the basis of data (machine reasoning, concerned with processes such as inference and optimisation) and robotics, or embodied AI, in which perceptual, reasoning, and action-related components are integrated in interaction with the physical environment (Sheikh et al., 2023). Crucially, the WRR report conceptualizes AI as a system technology rather than a discrete tool. Like general-purpose technologies such as electricity or the steam engine, AI exerts systemic influence across diverse domains, continuously improving while spawning complementary innovations. This framing situates AI within an interconnected infrastructure of algorithms, data, and hardware, emphasizing that intelligence emerges not from isolated computational entities but from complex socio-technical systems shaped by human oversight and institutional context (Sheikh et al., 2023).

In parallel, scholars in the communication and social sciences have developed definitions that highlight AI's interactive and cognitive dimensions. Gil De Zúñiga et al. (2024) define AI as the tangible capability of non-human machines to perform tasks, solve problems, communicate,

and act logically as biological humans do. This human-centred, interactional perspective emphasizes AI as a communicative and social agent, aligning with psychological approaches that study human responses to machine autonomy. Similarly, information-systems research defines AI as the capacity of machines to perform cognitive functions typically associated with human minds, such as perception, reasoning, learning, and creativity, without presupposing consciousness or moral understanding (Collins et al., 2021).

Empirical analyses of definitional practices reveal that most definitions are structured around five recurring criteria: human-likeness, learning ability, internal cognitive processes, goal-directedness, and problem complexity (Caluori, 2024). Among these, human-likeness and learning ability dominate, reflecting both the aspirational and adaptive aspects of AI as a field. The definitional discourse, therefore, mirrors broader philosophical and cultural questions about what constitutes "intelligence" and whether it can exist independently of human consciousness or emotion. In this sense, defining AI is as much a social and epistemic act as it is a technical one.

Synthesizing these perspectives, Artificial Intelligence (AI) can be understood as a class of computational systems that perceive and interpret their environment, learn and adapt from data, act with a degree of autonomy toward goal-directed objectives, and interact with humans or other systems through natural or symbolic interfaces. Such systems emulate selected cognitive functions of human intelligence (namely: reasoning, judgment, and communication) without possessing consciousness or moral experience. This definition situates AI within the domain of narrow AI, emphasizing specialized competence over general understanding. It also reflects a functional and psychological orientation: AI systems operate not only as computational instruments but as socio-cognitive actors capable of shaping human reasoning, trust, and moral evaluation.

In this sense, AI represents both a technological and cognitive extension of human decision-making. By combining rational optimization with adaptive learning, it integrates perception, reasoning, and action into self-improving processes that increasingly interact with social and ethical domains. Understanding AI in these terms establishes the conceptual groundwork for examining its role as an advisor in moral contexts: how algorithmic systems provide guidance, influence deliberation, and alter perceptions of responsibility and moral agency.

The subsequent passage dives further into Artificial Intelligence applications across various domains, highlighting its wide use in modern times.

2.2.3 AI in Different Application Fields

Artificial Intelligence (AI) has become one of the most transformative technologies of the twenty-first century, reshaping industries by enhancing decision-making, optimizing operations, and automating complex cognitive and procedural tasks. At its core, AI enables the analysis of vast and heterogeneous data streams, the detection of patterns, and the generation of predictive insights that augment rather than replace human expertise. Across economic and societal sectors, AI has evolved from an operational tool to a strategic co-decision-maker, increasingly embedded within professional routines that demand accuracy, efficiency, and adaptability (Berman et al., 2023; V. Kumar, 2024; Weng et al., 2024).

Healthcare

The healthcare sector has been among the earliest and most intensive adopters of AI, holding the potential for a fundamental transformation in the practice of medicine and the delivery of healthcare (Bajwa et al., 2021). Deep learning architectures now achieve diagnostic performance comparable to, and frequently surpassing, that of experienced radiologists in detecting pathologies such as malignant tumors, tuberculosis, and neurological disorders from radiographic, MRI, and CT images (Weng et al., 2024). Beyond diagnostics, AI accelerates drug discovery and development by mining chemical and biological databases for molecular candidates and by simulating clinical trial outcomes to optimize both time and cost (Weng et al., 2024).

Recent developments demonstrate that large language and multimodal models such as Med-PaLM M extend these capabilities by integrating textual, imaging, and genomic data to support clinical question answering, documentation, and medical report generation (Qu et al., 2025). In personalized medicine, machine learning systems synthesize genomic, phenotypic, and lifestyle data to recommend individualized treatment plans, particularly in oncology, where therapeutic success depends on genetic precision. AI-powered robotic surgery further enhances these advances by improving procedural accuracy, minimizing operative risk, and

accelerating postoperative recovery (V. Kumar, 2024).

At the patient interface, AI also supports mobile health and wellness applications that provide real-time feedback and personalized recommendations for prevention, rehabilitation, and chronic-disease management (Q. Liu et al., 2024; Qu et al., 2025). These systems increasingly operate in hybrid environments, combining cloud-based intelligence with privacy-preserving, on-device processing to enable secure, delay-sensitive decision-making at the point of care (Qu et al., 2025).

Importantly, AI technologies in healthcare function primarily as collaborative decision aids rather than as replacements for clinical expertise, augmenting human reasoning through evidence aggregation and probabilistic inference (Gaube et al., 2021, 2023). By enhancing diagnostic consistency and transparency, they support more efficient, accountable, and data-driven medical decisions while simultaneously raising new ethical challenges concerning fairness, bias, and responsibility in algorithmic health judgments (Q. Zhang et al., 2024).

Finance and Economics

In the financial domain, Artificial Intelligence has redefined decision-making under uncertainty through automation, prediction, and adaptive risk management. Algorithmic trading platforms employ reinforcement learning to detect short-term market dynamics and execute high-frequency trades informed by price fluctuations, trading volumes, and sentiment data (Weng et al., 2024). Machine learning–based robo-advisors provide individualized portfolio recommendations and continuously rebalance asset allocations to optimize the risk–return ratio in real time. Recent advances in large language and multimodal models have further expanded financial analytics by integrating heterogeneous data streams across text, numerical, and visual informations into unified decision-support frameworks capable of market summarization, trend forecasting, and sentiment extraction (Q. Liu et al., 2024; Qu et al., 2025).

Trust and acceptance remain critical determinants of AI adoption in finance. Empirical research demonstrates that investors' reliance on algorithmic recommendations depends strongly on perceived transparency, competence, and user focus: individuals prefer human advisors when decisions carry high personal or emotional involvement, whereas algorithmic advice is favored when it is seen as objective, data-driven, and impartial (Northey et al.,

2022). This duality highlights the social dimension of AI-mediated financial decisions, where confidence in computational neutrality must be balanced against the need for interpersonal understanding and accountability.

AI also supports financial integrity and regulatory compliance. Machine learning algorithms detect fraudulent activity and anti–money laundering (AML) violations by identifying anomalous transaction patterns in real time, while credit scoring models increasingly incorporate alternative data, such as behavioral, occupational, or consumption histories to extend financial inclusion and refine risk assessment (V. Kumar, 2024). However, these developments introduce new challenges in fairness and explainability. Studies reveal that automated financial and occupational classification systems may reproduce or amplify structural biases when trained on unbalanced datasets, underscoring the trade-off between predictive accuracy and equitable treatment (Q. Zhang et al., 2024).

Manufacturing and Industrial Engineering

Manufacturing industries increasingly deploy AI across all operational layers (from predictive maintenance to quality control and logistics optimization). Sensor-based monitoring systems apply machine learning to detect patterns in vibration, temperature, and pressure data, predicting component failures and enabling proactive maintenance scheduling (Weng et al., 2024). Computer-vision algorithms ensure production precision by identifying micro-defects and dimensional deviations, while adaptive process-control models maintain stability in key parameters such as pressure and material composition. Recent advances in mobile-edge and robotic intelligence extend these applications to fully autonomous manufacturing systems, where AI coordinates machinery, material flow, and quality inspection in real time (Q. Liu et al., 2024; Qu et al., 2025). At the managerial level, AI-driven supply-chain analytics forecast demand and optimize routing, reducing waste and operational downtime (V. Kumar, 2024). Through its integration into cyber-physical production environments, AI functions simultaneously as a control mechanism and a strategic decision-support tool, enhancing efficiency, flexibility, and system resilience.

Retail and Consumer Analytics

In the retail sector, AI drives personalization, demand forecasting, and revenue optimization. Recommendation systems analyze consumer data to generate personalized product suggestions that increase engagement and conversion rates (Weng et al., 2024). Conversational agents and chatbots based on natural language processing (NLP) enhance customer experience through 24/7 service interaction. Dynamic pricing algorithms continually adjust prices in response to fluctuations in demand, competitor activity, and inventory levels, a process now standard in e-commerce platforms such as Amazon and Netflix. Meanwhile, predictive demand models support strategic inventory management by accurately forecasting seasonal or regional trends, reducing stockouts and overproduction. These applications exemplify AI's role not only as a predictive engine but as a behavioral interface, that learns from user interaction to shape both commercial and experiential outcomes.

Transportation and Infrastructure

Transportation systems and urban mobility increasingly rely on AI for safety, efficiency, and environmental optimization. Autonomous driving technologies use sensor fusion, computer vision, and reinforcement learning to interpret dynamic road environments and execute navigation decisions in real time (V. Kumar, 2024). At the infrastructural level, deep reinforcement learning models regulate adaptive traffic signals, manage congestion, and optimize public transport scheduling (Weng et al., 2024). Predictive analytics in railway and air transport improve maintenance cycles and route planning, reducing operational costs and delays. These applications illustrate how AI transforms physical infrastructure into adaptive, data-driven ecosystems capable of self-optimization and continuous improvement.

Defense, Security, and Cyber Protection

AI's analytical and predictive capabilities have become central to modern defense and security operations. Autonomous reconnaissance systems and drones analyze aerial and satellite imagery to identify potential threats, while AI-based early-warning systems assist in real-time strategic decision-making (V. Kumar, 2024). In cybersecurity, anomaly-detection algorithms identify malicious behavior before breaches occur, automating the defense cycle. These tools

extend situational awareness and reduce response time, allowing defense actors to act with greater precision and reliability.

Scientific Research and Knowledge Generation

Scientific discovery increasingly depends on AI to process and interpret complex, high-dimensional datasets. Across disciplines machine learning supports the detection of new particles, prediction of molecular interactions, and mapping of neural activity with precision unattainable by manual analysis (Berman et al., 2023). In astronomy, AI systems classify celestial objects and forecast cosmic events with unprecedented accuracy, while multimodal and large-language models facilitate cross-domain integration by linking textual, numerical, and visual data sources (Q. Liu et al., 2024; Qu et al., 2025). Generative AI extends these capabilities by producing synthetic data, visualizations, and textual materials that accelerate hypothesis testing, experimental design, and knowledge dissemination (Sengar et al., 2024). Collectively, these technologies not only augment scientific productivity but also transform epistemic processes, automating parts of the research cycle previously constrained by human cognitive and temporal limits.

Emerging and Specialized Domains

Beyond the primary industrial sectors, AI is rapidly permeating niche applications. In engineering, deep learning improves structural diagnostics such as asphalt-pavement evaluation or distress detection using enhanced YOLO algorithms. In aerospace, generative models predict unmanned aerial system (UAS) trajectories and optimize flight safety. AI-driven cybersecurity tools safeguard digital assets through real-time anomaly detection and intrusion prevention. In education and software development, AI supports adaptive learning environments, code generation, and augmented reality (AR) interfaces that enhance remote collaboration (Weng et al., 2024). Collectively, these specialized implementations demonstrate the scalability and generalizability of AI technologies across both physical and digital domains.

Everyday Applications of Large Language Models

Large language models (LLMs) such as OpenAI's ChatGPT (OpenAI, 2024), Google's Gemini (Google DeepMind, 2023), Anthropic's Claude (Anthropic, 2023), and X's Grok (xAI, 2023) have rapidly transitioned from experimental systems to everyday cognitive companions. Based on transformer architectures, these models generate coherent, contextually adaptive text and multimodal outputs, enabling human-like interaction across communication, learning, and creativity. Their diffusion marks a qualitative shift from AI as a domain-specific analytical tool to a ubiquitous, personalized assistant integrated into routine digital environments (Q. Liu et al., 2024; Qu et al., 2025).

LLMs support a broad range of everyday and professional tasks. Users rely on them for writing, translation, coding, data analysis and for generating ideas, plans, or creative content. Integrated into productivity ecosystems such as Microsoft's Copilot or Google DeepMind (2023)'s Workspace AI, these systems enhance efficiency by providing real-time feedback and contextual recommendations (Wang et al., 2024). Mobile and on-device versions like Gemini Nano extend these capabilities through privacy-preserving, low-latency operation, bringing generative and decision-support functions directly to smartphones and personal computers (Qu et al., 2025).

Beyond professional utility, LLMs are increasingly embedded in informal learning and self-development contexts. They function as adaptive tutors that explain concepts, scaffold problem-solving, and enable self-directed or lifelong learning integrated into daily routines (Terzimehić et al., 2025). In parallel, their conversational nature facilitates social interaction, emotional support, and creative expression, illustrating how generative systems can mediate not only cognitive but also affective experiences online.

As these models become embedded in decision contexts, offering recommendations on various matters, they also raise issues of trust, fairness, and responsibility. Research highlights that users perceive LLMs as credible and socially competent sources of guidance, yet algorithmic persuasion and cognitive anchoring may shape user judgment in subtle ways (Northey et al., 2022; Q. Zhang et al., 2024). Understanding such mechanisms is crucial, as the ubiquity of conversational AI signals a broader transformation: from algorithmic tools that compute to socio-cognitive agents that advise, influence, and increasingly participate in moral reasoning

in everyday life.

Decision Support and Advisory Systems

Across all sectors, the unifying function of AI lies in its decision-support capacity. Whether diagnosing medical conditions, evaluating financial risks, or optimizing logistical networks, AI systems serve as second-layer cognitive instruments that assist rather than replace human agency. Decision-support algorithms synthesize complex information, generate probabilistic forecasts, and present interpretable insights that guide human judgment (Berman et al., 2023; Northey et al., 2022; Weng et al., 2024). In medicine, they enable clinicians to integrate diagnostic evidence and patient data for better outcomes (Gaube et al., 2021, 2023); in public policy, they inform evidence-based planning and resource allocation (V. Kumar, 2024). AI thus functions as an analytical co-agent, bridging computation and cognition to improve the accuracy, consistency, and accountability of complex decisions.

These application fields do not represent an exhaustive list but rather illustrate the broad scope and profound impact of AI across diverse domains. They serve as exemplars of AI's far-reaching influence, demonstrating its integration into and transformation of various professional and societal sectors.

The next chapter outlines the significant role of AI as a qualified advisor, explores both its acceptance and skepticism, and places particular emphasis on artificial intelligence in moral reasoning and its potential as a moral advisor.

2.2.4 AI as an Advisor

The concept of advice taking has been extensively studied in social and organizational psychology, where it is defined as the process of receiving, evaluating, and potentially integrating external input into one's own judgment (Bonaccio & Dalal, 2006). Integrating external advice generally enhances decision accuracy, although the extent of this benefit depends on situational factors such as task difficulty, individual confidence, and contextual stakes (Gino & Moore, 2007; Løhre & Halkjelsvik, 2024; Rader et al., 2017). As Artificial Intelligence evolves from a computational tool for data analysis into an autonomous advisor, it increasingly supports, guides, and shapes human decision-making. AI systems now occupy advisory roles once

reserved for humans, influencing not only the quality of judgments but also the confidence with which they are made. By providing guidance, predictions, and recommendations, AI functions not merely as a technical instrument but as a social advisor, that raises central psychological questions about when and why people follow AI advice, how they evaluate its credibility, and under which conditions they choose to accept or reject it.

Algorithm Aversion

Algorithm aversion describes the reluctance of individuals to accept, rely on, or act upon AI advice, even when algorithms demonstrably outperform human judgment (Castelo et al., 2019; Jussupow et al., 2020). The phenomenon represents a systematic deviation from rational choice, as rejecting a superior decision aid leads to lower accuracy and utility (Dietvorst et al., 2015). Rather than reflecting simple technophobia or ignorance, research increasingly shows that algorithm aversion is a structured psychological response rooted in expectations of competence, tolerance for imperfection, and perceived task fit (Jussupow et al., 2020; Mahmud et al., 2022).

Error Intolerance and Learning Expectations The most influential explanation for algorithm aversion concerns people's asymmetric tolerance for error. Across a series of forecasting experiments, Dietvorst et al. (2015) demonstrated that individuals who observe an algorithm make even a minor mistake subsequently lose trust in it and prefer to rely on inferior human forecasters. Although both human and algorithmic predictors err at similar rates, mistakes made by machines are judged more harshly because they violate expectations of mechanical perfection. Participants who had not witnessed algorithmic failure, by contrast, exhibited no such aversion, highlighting that the phenomenon is triggered specifically by observed imperfection (Dietvorst et al., 2015).

Algorithm aversion is not a baseline tendency but a reaction to unmet expectations of learning. When participants interacted with an algorithm in an impersonal, objective decision task, reliance was initially comparable to that on a human advisor. However, once the algorithm erred, trust declined sharply, unless participants were informed that the system could improve through learning. When the algorithm's ability to adapt was made explicit,

users became more tolerant of errors and continued to rely on its advice (Berger et al., 2021) This suggests that algorithm aversion arises less from general distrust and more from disappointment in the algorithm's perceived incapacity to recover or learn from mistakes.

Perceived Competence and Framing Effects A second explanatory mechanism concerns perceived competence, or expert power, of the advising agent. Hou and Jung (2021) demonstrated that the apparent contradiction between algorithm aversion and algorithm appreciation can be reconciled by framing differences in perceived expertise. Across three decision-making experiments, manipulating how the human and AI agents were introduced was sufficient to reverse results. When the human advisor was framed as a highly qualified expert and the algorithm as generic, participants preferred the human source. When the algorithm was framed as data-driven and superior, the pattern reversed. The strength of this framing effect shows that algorithm aversion and appreciation are not fixed attitudes but outcomes of contextual comparison between perceived competences.

The same logic appears in meta-analytic work (Jussupow et al., 2020). Aversion systematically increases when the human agent is portrayed as having high expertise or social closeness to the user, and decreases when the algorithm is presented as accurate, transparent, or human-assisted. Similarly, "algorithm factors" (e.g., accuracy, explainability, presentation style) and "individual factors" (e.g., anxiety, self-efficacy, familiarity) as primary drivers of aversion across 80 reviewed studies (Mahmud et al., 2022).Their synthesis shows that aversion arises when technical properties of the system (e.g. opacity or perceived rigidity) conflict with users' psychological needs for understanding, control, and security, whereas familiarity, transparency, and demonstrated learning capacity substantially mitigate reluctance to rely on AI advice. People evaluate AI advisors not in isolation but relative to perceived human alternatives. When algorithms are described as competent, transparent, and capable, aversion can dissipate or even reverse.

Task Characteristics and Affective Fit Beyond framing and competence, task features play a decisive role in shaping aversion. People are less willing to follow AI advice for tasks perceived as subjective or affect-laden than for those perceived as objective or analytic. Castelo et al. (2019) systematically manipulated task type across six experiments and found that

participants trusted algorithms less for subjective judgments (e.g., predicting joke funniness or recommending romantic partners) than for objective, quantifiable tasks (e.g., data analysis or route planning). The primary belief driving this effect is that algorithms lack affective or "human-nature" abilities such as empathy, intuition, and contextual understanding. When the same tasks were reframed to highlight their data-driven or quantitative nature, aversion decreased. Likewise, demonstrating that an algorithm could perform affective tasks (e.g., composing music or understanding emotion) eliminated the negative effect of task subjectivity on algorithm use. These results underscore that algorithm aversion is grounded in perceived mismatch between the advisor's capabilities and the task's demands.

Moral and Ethical Boundaries A particularly robust form of algorithm aversion arises in moral or ethically consequential decision domains. People are strongly averse to autonomous machines making moral decisions in domains such as medicine, law, and the military (Bigman & Gray, 2018). This aversion persists regardless of whether AI outcomes are objectively positive or negative and is mediated by perceptions of mind completeness. Participants attribute to machines some degree of agency but almost no experience, understood as the ability to feel or empathize. Because moral judgment is perceived to require both agency and experience, machines are regarded as incomplete moral agents. Consequently, people prefer to limit algorithms to an advisory role, reserving final responsibility for humans. Interestingly, increasing the algorithm's perceived expertise can reduce aversion slightly, but increasing its perceived emotional capacity does not, suggesting that moral legitimacy, not empathy simulation, defines the boundary of acceptable automation.

Applied Evidence Across Domains Empirical evidence across practical fields corroborates these mechanisms and reveals how context-specific expectations modulate algorithm aversion. In healthcare, individuals show resistance to AI-based diagnostic tools even when informed that such systems outperform human physicians. They report concerns that algorithms cannot account for personal characteristics or emotional nuance, leading to reduced trust and uptake (Longoni et al., 2019). In financial decision-making, investors express stronger belief in and intention to follow advice from human rather than algorithmic financial advisors, especially under conditions of high involvement. This effect is mediated by perceived information

credibility and the institution's perceived "customer focus," demonstrating that relational expectations, not computational ability, drive aversion (Northey et al., 2022). Similar findings emerge in service and consumer settings. Participants react negatively to automated decision-making when they perceive threats to autonomy, fairness, or transparency (Mohr & Kühl, 2021; Vitezić & Perić, 2021; Yalcin et al., 2022). These concerns reflect not simple resistance to technology but sensitivity to dignity and procedural justice. Qualitative work further illustrates that individuals justify algorithm aversion as a safeguard against dehumanization rather than as irrational technophobia: users cite loss of empathy, accountability, and contextual understanding as reasons for preferring human input (Mun et al., 2025). Across sectors, the same psychological mechanisms (expectations of competence, empathy, and moral responsibility) recur, manifesting differently depending on stakes and perceived social proximity.

Integrative Framework and Summary Synthesizing these findings reveals that algorithm aversion is best understood as a form of context-dependent trust calibration rather than a universal bias. People withdraw reliance when algorithms violate expectations of perfection, lack perceived learning ability, or appear mismatched to the affective or moral demands of a task. The strength and direction of aversion depend on perceived comparative expertise, framing, and the degree of human-likeness expected in the decision context. Meta-analyses emphasize that task characteristics, individual traits, and broader societal norms interact to determine acceptance levels (Jussupow et al., 2020; Mahmud et al., 2022). Algorithm aversion arises from rational social cognition: people reserve trust for agents they perceive as competent, adaptive, and morally legitimate. Understanding this dynamic is essential for designing AI systems that function effectively as advisors, particularly in moral and socially sensitive domains where human values and accountability remain indispensable.

Algorithm Appreciation

Algorithm appreciation describes the tendency for individuals to prefer, trust, or rely more on AI advice and decisions than on human input, even when both provide comparable information. It represents the mirror phenomenon of algorithm aversion, reflecting a positive bias toward

AI performance and impartiality. Rather than being driven by automation enthusiasm alone, Algorithm appreciation emerges when algorithms are perceived as more accurate, objective, or competent than human advisors. Across several decision-making contexts (like finance, healthcare, moral and everyday domains) research consistently finds that individuals sometimes place greater weight on AI advice than on human guidance (Kaufmann et al., 2023; Logg et al., 2019; You et al., 2022).

Cognitive and Trust-Based Mechanisms At the cognitive level, algorithm appreciation is primarily explained by trust in perceived analytical precision and reliability. Individuals tend to view AI agents as rational, data-driven, and immune to emotional or ideological bias. Logg et al. (2019) provided the first robust empirical evidence for algorithm appreciation, showing that participants weighted AI advice more heavily than human advice across several estimation and forecasting tasks. This preference persisted even when participants could not evaluate the algorithm's exact performance, suggesting a generalized belief in AI superiority.

Individuals exhibit stronger reliance on AI advice than on human advice, primarily due to higher trust in the algorithm's predictive accuracy. Interestingly, appreciation persisted when performance transparency was absent or presented only in aggregated form, but declined when detailed performance data were provided, which increased cognitive load. Individuals with high need for cognition continued to show appreciation even under high-load conditions, suggesting that cognitive engagement moderates the relationship between transparency and trust (You et al., 2022). Similarly, the precision of numerical information (such as presenting results as 74.975% rather than 75%) significantly increased behavioral intention to follow AI advice. This "preciseness effect" strengthens trust by reinforcing perceptions of analytic rigor and competence (J. Kim et al., 2021).

Trust emerges as a central mediator in the adoption of algorithmic recommendations across domains. Kelly et al. (2023) conducted a large-scale systematic review of AI acceptance and found that trust and perceived usefulness were the strongest predictors of behavioral intention across industries. In smart healthcare, trust even outweighed perceived usefulness as a predictor of willingness to adopt AI systems. Also, transparency, interpretability, and fairness perceptions significantly enhance trust in AI decision-makers. Participants considered interpretable AI systems that clearly explained their training processes more reliable and fair

than both opaque algorithms and human decision-makers, underscoring trust as the cognitive foundation of appreciation (Ashoori & Weisz, 2019).

Framing, Credibility, and Social Influence Algorithm appreciation depends strongly on how AI systems are introduced and socially positioned. When an AI advisor is framed as a highly competent domain expert, people are significantly more likely to prefer its advice over that of a human counterpart, indicating that perceived competence can reverse patterns of aversion into appreciation (Hou & Jung, 2021). This effect extends beyond textual framing to institutional credibility and social endorsement. Individuals are more willing to follow algorithmic recommendations when these systems are presented as credible, widely adopted, or backed by reputable organizations, as such cues signal trustworthiness and reduce the cognitive effort associated with independent evaluation (Alexander et al., 2018). Social proof thereby transforms technological novelty into normative legitimacy. At the same time, appreciation is often reinforced by comparative trust dynamics: when people perceive human institutions as biased, inefficient, or corrupt, they are more inclined to regard algorithms as objective and impartial alternatives (Gerlich, 2023). This dynamic aligns with the "machine heuristic," according to which AI systems are viewed as more neutral and fact-based than humans, particularly in domains where moral or ideological bias is salient (Araujo et al., 2020). Algorithm appreciation arises when competence, credibility, and impartiality converge, either by amplifying trust in the algorithm or by displacing trust from human decision-makers.

Task Type, Context, and Performance Conditions Beyond framing, appreciation varies systematically across task types and decision contexts. It is most pronounced in analytical or utilitarian domains, where precision, consistency, and error minimization are paramount. Across experimental paradigms, appreciation reliably occurs when tasks emphasize objective computation rather than affective or intuitive judgment (Kaufmann et al., 2023). Consumers, for instance, prefer algorithmic recommenders for utilitarian products that promise functional benefits, but revert to human input when choices involve hedonic or emotionally laden attributes (Yalcin et al., 2022). In medical contexts, trust in AI systems is highest for image-based specialties such as radiology and dermatology, which are perceived as data-intensive and low in moral risk, while more invasive applications such as robotic surgery elicit greater caution

(Yakar et al., 2022). Non-task experts similarly show stronger reliance when algorithmic advice is accompanied by clear, case-specific explanations and when they personally hold positive attitudes toward AI or possess higher AI literacy (Gaube et al., 2023). Comparable mechanisms apply in practical, non-clinical settings: farmers' willingness to adopt self-learning agricultural systems increases with perceived ease of use, usefulness, and controllability (Mohr & Kühl, 2021). Moreover, performance transparency moderates appreciation in nuanced ways: aggregated or concise performance feedback maintains trust, whereas overly detailed and cognitively demanding information can reduce it. Individuals high in need for cognition remain more appreciative under these conditions (You et al., 2022). These patterns indicate that appreciation peaks when task features align with AI's perceived analytical competence and when transparency, usability, and cognitive manageability foster sustained trust.

Moral and Value-Based Appreciation Algorithm appreciation also extends into moral and value-laden domains, where trust depends on perceived societal and ethical benefits. Public evaluations of AI applications are largely predicted by perceived moral qualities such as benefit, honesty, and low risk, with familiarity and prior experience further amplifying acceptance (Eriksson et al., 2025). Systems that visibly improve welfare, reduce drudgery, or perform prosocial functions are evaluated as both effective and morally legitimate. Fairness perceptions play a crucial role in this process: when algorithms are seen as transparent and unbiased, users attribute to them a form of procedural moral authority. Transparent model explanations and evidence of human oversight increase perceived fairness and trustworthiness, particularly in high-stakes decisions (Ashoori & Weisz, 2019). Likewise, across justice, health, and media contexts, AI decisions are often judged as equally or more useful and fair than those of human experts when impartiality is salient (Araujo et al., 2020). These results indicate that appreciation is not confined to utilitarian performance, it can also reflect moral endorsement when AI systems embody fairness, accountability, and benefit to others.

Integrative Synthesis and Theoretical Implications

Across the reviewed evidence, algorithm appreciation consistently emerges when three conditions co-occur: the system is perceived as competent and credible, the task highlights

objective accuracy, and the decision process appears transparent and fair. These features trigger a calibration of trust in which people selectively delegate evaluative authority to algorithms that fulfill expectations of expertise and impartiality. Meta-analyses confirm that both algorithmic properties (accuracy, explainability, interactivity) and individual dispositions (trust, familiarity, self-efficacy) jointly determine the acceptance threshold; transparency and demonstrated learning capability further mitigate residual hesitation (Mahmud et al., 2022). Algorithm appreciation and aversion therefore represent two poles of a single trust-regulation process: appreciation arises when technical and moral expectations are met, aversion when they are violated (Hou & Jung, 2021). Understanding this dynamic is critical for assessing AI as an advisory system, since appreciation entails not only confidence in computational precision but also the willingness to extend moral legitimacy to algorithmic judgment.

Performance and Accuracy of AI Advisors

AI systems often match or surpass human performance in both objective and subjective decision domains. Recommender systems can more accurately predict individual preferences than human acquaintances, demonstrating superior accuracy even for affect-laden judgments such as humor appreciation (Yeomans et al., 2019). In social advisory contexts, AI-generated guidance is perceived as more balanced, empathetic, and complete than that of professional human columnists, even when response length and style are controlled, indicating that perceived quality advantages persist when superficial cues are eliminated (Howe et al., 2023). Similar findings appear in quantitative and analytic domains, where individuals rely more on algorithmic than on crowd-derived advice, particularly as task difficulty increases, implying that cognitive load amplifies appreciation of AI expertise (Bogert et al., 2021). Moreover, when machine-learning models automatically discover optimal decision strategies and express them as clear procedural instructions, human decision quality improves substantially in realistic tasks such as mortgage selection or travel planning (Becker et al., 2022).Comparable effects emerge in medical contexts, where non-expert decision-makers achieve significantly higher diagnostic accuracy in X-ray tumor detection when receiving correct and explainable AI advice, demonstrating that transparent AI support can enhance expert-level reasoning across domains (Gaube et al., 2021). These results converge on the insight that AI advice can enhance human

reasoning not merely by producing accurate outcomes but by rendering complex strategies cognitively accessible.

Dynamics of Reliance and Trust Calibration

Reliance on AI advisors is dynamic and feedback-sensitive rather than static. Human confidence in AI evolves with performance outcomes: positive feedback about successful AI predictions significantly strengthens trust in the system, while positive feedback about one's own success increases self-confidence and may reduce future reliance (Chong et al., 2022). Early algorithmic errors can have disproportionate long-term effects, reducing subsequent adherence to advice even when later accuracy improves, showing that initial impressions shape trust trajectories (Gill et al., 2024). When decision-makers receive a second opinion from either another AI or a human peer, they tend to reduce overreliance but risk greater underreliance unless given discretion over when to solicit additional input (Lu et al., 2024). These studies demonstrate that AI trust follows an adaptive calibration process: users continually update reliance based on observed performance, feedback valence, and perceived control over the advisory relationship.

Contextual Fit Between Advisor and Decision-Maker

The effectiveness of AI advice depends on how well the advisor's competencies fit the decision environment and the decision-maker's cognitive style. Utilization increases when the AI's expertise complements, rather than mirrors, the user's skills and when the task is perceived as objective or data-driven (Baines et al., 2024). Individuals with analytical decision preferences display higher trust and reliance on AI forecasts, especially in long-term or complex scenarios, whereas intuitive decision-makers with low digital trust favor human experts (Mayer et al., 2023). Perceived transparency and interpretability of the AI's reasoning strengthen satisfaction and willingness to use advice by helping users construct causal mental models of how recommendations are derived (Baines et al., 2024; Yeomans et al., 2019).

Mechanisms of Appropriate Reliance

Appropriate reliance emerges when users balance perceived competence with control and accountability. When AI expertise clearly exceeds human competence in a given domain, reliance increases proportionally, but only if the system's authority is transparent and aligned with user goals (Baines et al., 2024). Perceptions of complementarity (where the AI contributes distinct analytical strengths) foster cooperative decision-making, whereas perceptions of replacement provoke resistance. Experimental findings demonstrate that individuals adjust their weighting of AI advice dynamically. After successful integration, trust stabilizes at intermediate, calibrated levels that optimize accuracy without forfeiting autonomy (Chong et al., 2022; Gill et al., 2024). This pattern illustrates that effective human–AI advising is neither blind acceptance nor categorical rejection but a process of iterative trust negotiation.

Moral and Social Dimensions of AI Advice

AI advisors are increasingly used in domains involving moral judgment and social reasoning. Both human and robotic advisors are perceived as more trustworthy when recommending actions that favor collective welfare over individual benefit, suggesting that utilitarian alignment enhances perceived moral legitimacy (Starr et al., 2021). Philosophical and applied research proposes the development of artificial moral experts capable of dialogical engagement rather than prescriptive verdicts. Systems such as the hypothetical "SocrAI" are designed to help users clarify reasoning, identify inconsistencies, and preserve moral autonomy, thereby addressing concerns about overreliance and paternalism in algorithmic ethics (Rodríguez-López & Rueda, 2023). These approaches position AI not as a moral authority but as a reflective partner in ethical deliberation.

The growing potential of AI as a source of moral guidance raises a fundamental question about its position within the moral domain itself. The following section therefore moves beyond AI as an advisor and addresses its theoretical status in relation to artificial morality and moral decision-making.

2.2.5 Artificial Intelligence in Morality and Moral Decision-Making

Introduction: The Rise of Artificial Morality

Artificial intelligence (AI) increasingly operates within ethically charged domains (such as medicine, law, and autonomous vehicles), where its actions can carry moral consequences. This development has turned the question of whether machines can or should make moral decisions into one of the central challenges for both psychology and ethics (Bigman & Gray, 2018; Bonnefon et al., 2020; D. S. Kumar, 2025). While AI systems excel at analytical reasoning and outcome optimization, people remain reluctant to attribute them genuine moral authority, primarily because they lack experiential and emotional capacities considered essential to human moral cognition (Bigman & Gray, 2018; D. S. Kumar, 2025). This tension between computational rationality and moral intuition defines the contemporary problem of artificial morality, the difficulty of reconciling algorithmic decision logic with the affective, social, and contextual foundations of moral judgment (Bonnefon et al., 2020).

Empirical research reveals that people apply distinct moral norms to artificial and human agents. Robots and AI systems are typically expected to act according to utilitarian principles, yet they are blamed more harshly when they fail to do so (Malle et al., 2015, 2016; Y. Zhang et al., 2023). This moral asymmetry reflects a deeper ambivalence: humans acknowledge AI's cognitive agency but deny it moral experience, viewing machines as capable of reasoning but incapable of feeling (Y. Liu & Moore, 2025; Z. Zhang et al., 2022). Consequently, moral evaluations of AI hinge not only on the results of its actions but also on the perceived authenticity of its moral reasoning, which depends on attributions of empathy, intention, and moral understanding (Bigman & Gray, 2018; Y. Liu & Moore, 2025).

At the same time, AI's involvement in moral contexts is no longer hypothetical. Autonomous systems already assist in life-and-death decision-making (e.g. medical triage, resource allocation, or autonomous driving) and require societies to confront questions of accountability, fairness, and moral trust (Hendrycks et al., 2023; Niforatos et al., 2020). These developments have broadened the philosophical debate from whether AI can make moral decisions to how humans and machines should collaborate in making them (Etienne, 2022; Formosa & Ryan, 2021). This shift has given rise to a new interdisciplinary research field concerned with Artificial Moral Agents and Artificial Moral Advisors (AMAs), two systems designed not

merely to automate ethical reasoning but to assist and augment human moral deliberation (Cervantes et al., 2020; Giubilini & Savulescu, 2018; Rodríguez-López & Rueda, 2023). AMAs vary in design: some aim for autonomous decision-making in rule-based environments, while others, such as dialogical "Socratic" systems, seek to foster moral reflection through interactive reasoning rather than prescriptive output (Rodríguez-López & Rueda, 2023).

In parallel, psychological and behavioral research explores how people perceive and evaluate such systems. Studies show that acceptance of moral AI is conditional on contextual fit and human oversight: people prefer AI assistance in analytical, data-driven contexts but reject it in subjective or value-laden domains where empathy and moral intuition are essential (D. S. Kumar, 2025; Z. Zhang et al., 2022). These findings indicate that moral legitimacy in artificial systems arises not from technical accuracy alone but from the degree to which they align with human moral expectations and social norms.

Against this background, the following sections examine (a) the ethical and normative frameworks that define the boundaries of artificial morality, (b) empirical research on how humans perceive AI as moral agents, and (c) emerging evidence on AI as an actual moral advisor.

The Ethical Architecture of Artificial Morality The ethical evaluation of AI in moral domains revolves around two persistent challenges: (a) the alignment problem and (b) the responsibility gap. The alignment problem refers to the difficulty of embedding complex, sometimes conflicting human values into algorithmic systems in a manner that remains context-sensitive, interpretable, and normatively legitimate (Hendrycks et al., 2023). Early efforts to formalize machine ethics, such as rule-based moral codices reminiscent of Asimov's "Three Laws of Robotics," demonstrated the limits of top-down moral architectures in capturing the nuance and context dependence of human ethical reasoning (Bonnefon et al., 2020). The responsibility gap, by contrast, arises when semi-autonomous or autonomous systems act in ways that yield moral consequences without a clear locus of accountability (Constantinescu et al., 2022; Shank et al., 2019). In such hybrid constellations, moral blame is often diffused between human supervisors and AI agents, creating conditions under which no actor appears fully answerable.

Delegation versus Augmentation Contemporary discourse distinguishes two normative pathways for integrating moral competence into machines. The first approach (moral delegation) seeks to endow AI with the capacity to make independent ethical judgments, developing Artificial Moral Agents (e.g. in selfdriving cars) (Cervantes et al., 2020; Formosa & Ryan, 2021). Proponents argue that delegation could promote consistency and timeliness in morally charged environments such as medical triage or autonomous driving. Yet critics contend that moral delegation reduces ethics to formal optimization, eroding human responsibility and practical wisdom (Van Wynsberghe & Robbins, 2019). Because algorithms lack phenomenal consciousness, empathy, and emotional understanding, their "moral" outputs remain simulations of deliberation rather than genuine moral acts (D. S. Kumar, 2025).

The second approach (moral augmentation) conceptualizes AI as a facilitator of moral reflection rather than a substitute for it. Giubilini and Savulescu (2018) introduced the notion of the Artificial Moral Advisor as a system that enhances human moral decision-making by counteracting biases and information constraints. Drawing on the "ideal observer" tradition, such advisors aim to be consistent, impartial, and aligned with users' own moral standards. Ethical augmentation thus aspires to extend moral deliberation while preserving human agency and accountability (Constantinescu et al., 2022; Etienne, 2022). This conception corresponds to current psychological models of human–AI complementarity, according to which artificial systems amplify analytic reasoning while humans provide contextual and affective grounding.

Accountability, Transparency, and Oversight Across theoretical and empirical studies, a stable consensus emerges: moral accountability must remain anchored in human agents. Surveys indicate that the overwhelming majority of respondents reject fully autonomous moral AI and favor hybrid arrangements in which humans retain final responsibility (D. S. Kumar, 2025). Transparency and interpretability are repeatedly identified as prerequisites for moral legitimacy, because users can only ascribe accountability when they understand the logic of an algorithmic recommendation or decision (Hendrycks et al., 2023; Sommaggio & Marchiori, 2020). Explainable AI (XAI) mechanisms therefore constitute not only a technical desideratum but a moral one, ensuring traceability of reasoning and preserving public trust.

Empirical findings further reveal that collaboration with AI alters the distribution of moral blame. When humans oversee or confirm algorithmic decisions, they tend to be judged less

culpable for ensuing harms than actors operating independently. This describes a phenomenon of responsibility diffusion within human–AI teams (Shank et al., 2019). Meanwhile, AI agents themselves are often perceived as causally responsible for outcomes yet rarely attributed forward-looking moral responsibility, reflecting a deep reluctance to equate cognitive competence with moral accountability (Lima et al., 2021). This asymmetry underscores the distinction between intelligence and intentionality: while computational systems can emulate reasoning, they cannot instantiate the experiential consciousness that grounds human moral responsibility (Bigman & Gray, 2018; Bonnefon et al., 2020).

Normative Orientation and Conceptual Boundaries Philosophical analyses converge on the view that AI should function as a moral instrument (in form of a tool that expands the scope and consistency of human moral reasoning, rather than as a moral subject in its own right) (Constantinescu et al., 2022; Etienne, 2022). Deontological or consequentialist architectures can mimic aspects of ethical reasoning but remain incapable of resolving contextual ambiguity or moral pluralism (Sommaggio & Marchiori, 2020). Consequently, ethical AI design must emphasize adaptability, transparency, and continuous human supervision.

Three normative principles demarcate the conceptual boundaries of artificial morality. First, human accountability is non-delegable: final moral responsibility must always reside with human agents. Second, transparency and interpretability are indispensable for sustaining moral legitimacy and societal trust. Third, moral augmentation over moral substitution represents the ethically sustainable trajectory of AI development, ensuring that machines assist rather than replace human deliberation.

These principles establish the ethical architecture of artificial morality and frame the empirical question of how humans perceive and interact with AI systems in moral contexts. The next section explores how AMAs are perceived.

Human Perception of AI in Moral Contexts

Understanding how humans perceive artificial systems in moral situations is crucial for evaluating their ethical legitimacy. Empirical research across psychology, philosophy, and human–robot interaction consistently demonstrates that people attribute to AI a form of

partial moral agency, acknowledging its intelligence and autonomy while denying it emotional and experiential depth (Bigman & Gray, 2018; Y. Liu & Moore, 2025; Z. Zhang et al., 2022). This ambivalence produces asymmetrical patterns of moral judgment, where AI is expected to reason impartially but is still denied full moral standing when its behavior violates human moral expectations.

Agency Without Experience: The Incomplete Moral Mind Humans perceive AI as capable of deliberate action (agency) but devoid of subjective experience (sentience) (the ability to feel and empathize) (Bigman & Gray, 2018). This cognitive schema leads observers to view AI as rational but morally incomplete: competent in logic yet lacking the warmth required for moral legitimacy (Z. Zhang et al., 2022). Consequently, in morally charged domains such as medicine, military operations, or autonomous driving, people prefer human decision-makers, associating human emotions with empathy and moral concern (D. S. Kumar, 2025). This dichotomy defines the "incomplete moral mind" problem where AI systems are granted instrumental trust but denied the affective foundation of moral authority.

Moral Asymmetries in Judgment and Blame Experimental evidence reveals distinct asymmetries in how people evaluate the moral behavior of human versus artificial agents. In sacrificial dilemmas such as the Trolley Problem, participants expect robots and AI to act according to utilitarian principles, but condemn them more severely when they fail to do so (Malle et al., 2015, 2016; Y. Zhang et al., 2023). This pattern is called the moral human–robot asymmetry and it reflects the assumption that machines should maximize collective welfare precisely because they lack emotions. Conversely, when AI executes the utilitarian choice, its decision is judged as more acceptable or even morally superior relative to the same action performed by a human agent (Z. Zhang et al., 2022).

Further studies demonstrate that physical embodiment and social cues moderate this effect. Humanoid robots, perceived as more humanlike, elicit greater praise for prosocial actions but also greater blame for transgressions (Malle et al., 2016). In contrast, purely mechanical or abstract AI systems are judged more leniently for identical outcomes, implying that anthropomorphic design activates moral schemas normally reserved for humans.

Contextual and Cognitive Determinants Moral judgments of AI depend strongly on context, task characteristics, and perceived domain appropriateness. People are more willing to accept AI decision-making in analytical or quantifiable domains (e.g. statistical prediction or risk assessment) than in subjective or emotionally laden contexts (such as end-of-life medicine or distributive justice) (D. S. Kumar, 2025; Z. Zhang et al., 2022). Studies of autonomous drones and defense systems show that participants apply similar utilitarian norms to both human and machine agents but differentiate blame attribution: humans are faulted for intention, while machines are faulted for design or oversight failures (Malle et al., 2019).

Likewise, people's comfort with machine ethics increases when moral choices are framed as procedural or rule-based rather than emotional. Experiments using immersive virtual moral dilemmas reveal that participants' acceptance of machine decision-making rises when scenarios emphasize objectivity, indicating that perceived fit between task type and agent competence shapes moral trust (Niforatos et al., 2020).

Belief Alignment and Cultural Moderators Acceptance of AI moral verdicts is further moderated by ideological and cultural factors. Research by Y. Liu and Moore (2025) shows that users' willingness to endorse AI moral judgments increases when the AI's stance aligns with their own politico-moral orientation. Perceived fairness and trust rise with belief alignment, while dissonance triggers moral resistance even when algorithmic reasoning is consistent. Cross-cultural analyses indicate that familiarity and technological exposure mitigate aversion to machine morality: individuals with higher AI literacy exhibit greater tolerance for utilitarian outcomes and less insistence on human oversight (Z. Zhang et al., 2022), highlighting that moral acceptance is socially constructed and dynamically shaped by ideology, experience, and education.

Integrative Perspective: Conditional Moral Acceptance Synthesizing these findings, human perception of moral AI can be understood as a dynamic equilibrium between rational approval and affective hesitation. People attribute cognitive agency to AI when it performs consistently and transparently but withdraw moral approval when its actions conflict with human norms of empathy or fairness (Bigman & Gray, 2018; Malle et al., 2015; Z. Zhang et al., 2022). This conditional acceptance produces a pattern of calibrated trust: AI is expected to be impartial

and efficient, yet it is denied full legitimacy as a moral actor. Over time, increased exposure and anthropomorphic design may normalize limited moral agency in artificial systems, but current evidence suggests that moral trust in AI remains constrained by expectations of emotional understanding, contextual sensitivity, and shared human values (D. S. Kumar, 2025; Y. Liu & Moore, 2025).

This psychological ambivalence forms the basis for the next section, which addresses how AI systems not only act in moral contexts but increasingly advises humans, shifting from moral agents to moral counselors whose guidance can shape human moral decision-making.

Influence of AI Advice on Human Moral Decision-Making

The emergence of Artificial Moral Advisors (AMAs) and conversational AI systems, such as OpenAI (2024)´s ChatGPT, has made moral advice a practical rather than hypothetical phenomenon. Empirical evidence demonstrates that algorithmic advice substantially shapes human moral judgment, even when people are explicitly aware that it originates from an artificial source. This influence can both enhance and corrupt moral reasoning.

Susceptibility to Artificial Moral Advice Across multiple studies, participants exhibit a strong tendency to incorporate AI-generated advice into their moral judgments. Even when explicitly informed that the advisor is an algorithm, individuals systematically underestimate the degree to which their own decisions are influenced (Krügel et al., 2023a, 2023b). In controlled experiments, ChatGPT's moral recommendations significantly shifted users' choices toward the advised option, regardless of whether the advice included justification or moral reasoning (Krügel et al., 2023b). This suggests that people often treat algorithmic advice as an epistemic shortcut, a convenient way to "escape" the cognitive and emotional discomfort of moral deliberation. Transparency regarding the artificial source does little to mitigate this effect (Krügel et al., 2023a). Research showed that up until recently, moral advice from AI systems was highly inconsistent, with identical dilemmas eliciting opposing responses depending on their framing (Krügel et al., 2023b). More recent evidence suggests that most recent LLMs produce more stable and internally consistent moral judgments (Coleman et al., 2025). Nevertheless, despite recognizing earlier inconsistencies, participants tended

to attribute substantial authority to AI recommendations, suggesting that susceptibility to moral advice operates largely below the level of conscious evaluation. In line of evidence, Krügel et al. (2022) demonstrated that when humans make moral decisions jointly with an algorithm, they frequently rely on its judgment even in ethically questionable contexts. Rather than correcting flawed or unethical algorithmic advice, participants often used them as post-hoc justification for self-serving outcomes, thereby diffusing moral responsibility and weakening personal accountability. This finding illustrates how algorithmic collaboration can inadvertently promote moral disengagement. Humans outsource not only cognitive effort but also ethical deliberation to the machine. Extending these insights, Krügel et al. (2025) examined ChatGPT as a contemporary moral advisor and found that its inconsistent yet persuasive moral stances significantly shaped users' judgments, even when the algorithm's artificial nature was disclosed. The influence persisted irrespective of argument quality or justification, revealing that moral persuasion through AI operates largely on heuristic and affective rather than deliberative pathways.

Corruptive and Enabling Potentials The moral influence of AI is not uniformly benign. Empirical findings reveal that exposure to unethical AI advice can increase dishonest behavior to the same extent as equivalent human advice (Leib et al., 2021, 2023). Participants who received dishonesty-promoting recommendations from an algorithm were equally likely to cheat as those advised by human peers, even when the artificial origin was fully disclosed (Leib et al., 2023). This parity in influence implies that users grant moral weight to algorithmic counsel independent of perceived intentionality.

Complementary evidence demonstrates that AI advisors may erode moral responsibility by enabling moral disengagement. When algorithms act as partners or delegates in decision-making, humans often exploit this shared structure to displace guilt or accountability ("ethical blind spots")(Köbis et al., 2021; Krügel et al., 2023a). In joint human–AI decision-making scenarios, people tend to accept AI-driven unethical outcomes when they personally benefit, rationalizing that moral responsibility is shared or externalized. Experimental data show that humans supervising AI systems are blamed less for moral violations than solo decision-makers, confirming a "responsibility diffusion" effect in hybrid moral contexts (Shank et al., 2019).

Mechanisms of Moral Influence Several psychological mechanisms explain why AI moral advice is persuasive. First, algorithmic advice leverages the machine heuristic (the belief that algorithms are objective, data-driven, and free from human bias) (Araujo et al., 2020). This perceived impartiality amplifies trust, even when accuracy or moral coherence is low. Second, the cognitive offloading hypothesis suggests that individuals delegate moral reasoning to AI to reduce cognitive effort and emotional conflict, particularly in high-stakes or ambiguous dilemmas (Etienne, 2022; Krügel et al., 2023b). Third, feedback loops between human and AI agents reinforce reliance: once users observe that AI advice leads to acceptable or reward-consistent outcomes, they are more likely to defer to it again, forming habitual dependence patterns (Gill et al., 2024). However, these same mechanisms contribute to underdeveloped moral reflection. Excessive reliance on AI advice reduces users' engagement with their own ethical reasoning, gradually weakening their capacity for independent moral judgment (Rodríguez-López & Rueda, 2023).

Moderators and Boundary Conditions Not all advice contexts produce the same effects. The impact of AI moral advice depends on the ethical framework employed, the domain of the dilemma, and the user's cognitive style. Experiments comparing deontological, virtue-based, and Confucian role-ethics advice show that rule-based (deontological) framing is marginally more effective in promoting honesty than virtue or role-based appeals, although overall effects remain small (B. Kim et al., 2021). This suggests that prescriptive moral rules may resonate more with users when issued by an artificial system than abstract appeals to virtue or social harmony. Contextual factors such as perceived impartiality, domain familiarity, and moral load also moderate influence. Users with higher AI literacy or reflective moral orientation are somewhat more resistant to manipulative advice (Y. Liu et al., 2022). In contrast, individuals low in reflective reasoning or experiencing moral fatigue show stronger conformity effects, following AI advice even when inconsistent or morally dubious (Krügel et al., 2023b). This indicates that cognitive depletion and trust calibration jointly determine whether AI functions as a moral amplifier or a moral hazard.

Integrative Perspective: Moral Calibration in Human–AI Interaction The empirical literature converges on a model of moral calibration between humans and AI. Algorithmic

advice shapes moral decisions by modulating cognitive load, perceived legitimacy, and social responsibility attribution. When users treat AI as a reflective partner, moral consistency and deliberative awareness can improve (Etienne, 2022; Giubilini & Savulescu, 2018). However, when advice is followed passively or strategically, moral agency erodes, and responsibility diffuses (Köbis et al., 2021; Shank et al., 2019). The challenge, therefore, lies not in whether AI should participate in moral discourse but in how systems and contexts can be structured to preserve human accountability and ethical reflection while leveraging algorithmic insight.

Taken together, these findings demonstrate that AI moral advice is not a neutral input but a powerful social cue that reconfigures moral cognition. Its effects depend on both design transparency and user engagement. Both variables are central to understanding when Artificial Moral Advisors enhance or endanger moral decision-making.

2.3 Integration and Research Gaps

The reviewed literature converges on three central insights. First, AI has evolved from a passive analytical instrument into an active advisory system that can influence human judgment and decision-making. Second, acceptance of AI advice is highly context-dependent, oscillating between aversion and appreciation according to task complexity, perceived competence, trust calibration, and motivational factors. Third, while AI has been extensively examined in technical and managerial domains, systematic research on AI as a moral advisor remains limited, despite its growing presence in ethically sensitive contexts.

Open questions persist regarding how situational and psychological factors, such as dilemma difficulty, social proximity to affected individuals, and the framing or explainability of moral advice, shape adherence to AI advice. Addressing these questions requires integrating moral-psychological models of intuition and reasoning with theories of social influence and human–AI interaction. The preceding chapters outlined how moral judgment arises from the dynamic interplay between intuitive and deliberative processes, shaped by emotional, social, and contextual influences. Within this framework, external input such as advice represents an additional cognitive and social cue that can guide, correct, or bias moral decision-making. The dual-process perspective explains why advice can penetrate intuitive moral reasoning: it provides a cognitively efficient heuristic that reduces effort while maintaining sufficient

accuracy, yet its influence depends on perceived credibility, trustworthiness, and situational fit. Integrating insights from the heuristics-and-biases program, moral intuitionist theories, and research on advice taking suggests that AI systems now occupy a novel position in this cognitive landscape, acting simultaneously as rational information processors and as social agents that trigger intuitive acceptance cues. Empirical findings consistently demonstrate that advice from artificial systems can meaningfully shape human moral decision-making. Across a growing body of studies, AI and human advisors exert comparable influence on moral decisions, even when the AI source is portrayed as inconsistent, untrustworthy, or transparently artificial (Krügel et al., 2022, 2023a, 2023b, 2025; Leib et al., 2021, 2023). Transparency about the AI origin does not substantially reduce compliance, indicating that people treat AI advice as a persuasive cue rather than purely informational input. Nevertheless, the mechanisms that govern when and why individuals follow or resist such advice remain underspecified. Most existing paradigms rely on abstract, sacrificial dilemmas with limited emotional engagement, thereby restricting ecological validity and obscuring how AI influence unfolds in everyday moral contexts. The present research program addresses four interrelated gaps.

First, we yet have to come to a robust effect on across variations in dilemma type, complexity, and difficulty. The aim is to extend existing findings and demonstrate the stability of AI influence across dilemmas of varying difficulty, thereby contributing to a more comprehensive and empirically grounded understanding of this phenomenon, increasing relevance for future human–AI interaction in moral decision-making.

Second, the role of social proximity and emotional relevance in the reception of advice remains insufficiently examined in the context of moral decision-making. Previous work has paid limited attention to how altruistic versus egoistic response tendencies unfold across varying degrees of social proximity. Consequently, it is still unclear whether heightened affective involvement reduces reliance on artificial advisors or whether such dynamics generalise to more ecologically valid, everyday dilemmas.

Third, existing research offers little insight into how individuals respond when human and AI advisors provide convergent or conflicting advice. Most studies have analysed advisory sources separately, despite the fact that real-world decision-making commonly involves integrating input from multiple advisors whose positions may either converge or conflict. It therefore

remains open whether individuals exhibit systematic preferences or biases when weighing human and AI advice simultaneously, or whether both sources are treated on largely comparable terms.

Fourth, the dynamics of retrospective decision change and how individuals revise or reaffirm their initial moral decisions after receiving AI advice. Most moral decisions are made intuitively and rapidly, guided by affective heuristics rather than deliberate reasoning. In everyday contexts, people often arrive at an immediate moral judgment that serves as an anchor for subsequent reflection. It is essential to understand how AI advice interacts with these intuitive first impressions: whether it consolidates, shifts, or destabilizes them upon reconsideration. This approach captures the temporal dimension of moral decision-making and provides insight into the stability of AI influence once initial intuitions have been formed, thereby contributing to a more ecologically valid, process-oriented understanding of human–AI moral interaction.

These research gaps extend prior work by combining internal experimental control with improved external realism. They examine not only whether AI influences moral decision-making, but also how contextual and cognitive boundary conditions shape this influence. Addressing these gaps advances both moral psychology and human–AI interaction research by specifying the mechanisms through which Artificial Moral Advisors shape judgment and by delineating the limits of their persuasive power in real-world decision contexts.

2.3.1 Aim of the Dissertation

Building on the integrated theoretical and empirical foundation outlined in the preceding sections, the present dissertation seeks to systematically examine how and under which conditions artificial moral advice influences moral decision-making. Rather than treating advice as a unitary phenomenon, the research program adopts a multidimensional approach, investigating how the persuasive impact of AI advice varies as a function of contextual, social, and temporal features of moral decision-making.

First, the dissertation examines the stability of AI influence across different moral structures and levels of difficulty. By varying dilemma type, dilemma difficulty, and the presence of explicit justification, the studies assess whether artificial moral advice exerts a consistent effect

across different moral conflicts. This allows for an evaluation of whether advice functions as a generalizable heuristic across domains or whether its effectiveness depends on specific cognitive or normative properties of the situation.

Second, the research extends beyond abstract moral scenarios by incorporating socially embedded, emotionally relevant dilemmas. By manipulating social proximity and contrasting altruistic with egoistic behavioral options, the research investigates how relational closeness and affective involvement influences reliance on AI advice. In doing so, the research addresses the interpersonal dimension of moral decision-making and examines whether artificial advice remains influential when decisions affect psychologically close or distant others.

Third, the dissertation addresses that moral decision-making rarely occurs in the presence of a single, isolated advisor. Individuals are often confronted with multiple advisors whose advice may converge or conflict. To capture this dynamic, the research introduces conditions in which participants receive moral advice from both human and AI advisors, that either converges or conflicts. This design enables a direct comparison of how individuals weigh, integrate, or disregard competing advice and tests whether human advice is prioritized over AI advice under normative disagreement.

Fourth, the research extends beyond initial decision formation to examine the temporal dimension of moral judgment. By introducing advice after a judgment has already been made, this research assesses the extent to which AI advice modifies, destabilizes, or reinforces existing moral decisions. This retrospective perspective captures a psychologically more realistic process in which initial intuitions anchor subsequent reflection, allowing us to assess whether AI shapes not only initial decisions but also already-formed moral decisions.

These objectives provide a comprehensive, process-oriented investigation of artificial moral advice. They elucidate not only whether AI can influence moral decision-making, but how this influence unfolds across diverse contexts, relational structures, and temporal stages of evaluation in relation to human advice. In doing so, the dissertation contributes to a more nuanced understanding of human–AI interaction in ethically significant environments and offers an empirically grounded framework for anticipating the role of AMAs in future decision-making landscapes. The following chapter translates this theoretical framework into empirical research, introducing four empirical studies.

3 Study 1: Can AI Change our Morals? The Influence of Artificial Advice on Moral Decision-Making in Dilemmas with Varying Difficulty.

3.1 Abstract

In this study, we examined how the advisor and normative framing of moral advice influence decisions across moral dilemmas of varying difficulty. A total of $N = 1{,}168$ participants completed nine classical dilemmas (non-moral, impersonal moral, and personal high-conflict scenarios) in an online experiment. Participants were randomly assigned to receive deontological, utilitarian, or no advice, with the advice framed as originating either from AI or a human expert. Moral decisions and subjective ratings of ease, certainty, and responsibility were analyzed using mixed-effects models. Across all dilemma types, deontological advice reliably increased the likelihood of deontological decisions, whereas utilitarian advice exerted comparatively weak influence. The persuasive impact of advice did not differ systematically between AI and human advisors, indicating that advisor labels alone did not shape behavioral uptake. Advice modestly altered subjective experience: it increased perceived ease in high-conflict dilemmas and reduced perceived responsibility in less severe cases. Individual differences in wisdom and moral identity showed nuanced associations with advice-following and subjective appraisals but did not overturn the central pattern of normative asymmetry. These findings suggest that external moral advice functions primarily as a content-driven support mechanism in morally challenging situations. Deontological, norm-congruent advice is readily adopted, regardless of whether they are attributed to an AI or a human, whereas norm-deviant utilitarian counsel is comparatively resistant to uptake.

Keywords: moral decision-making, Artificial Intelligence, moral advice, dilemma difficulty

3.2 Introduction

This study investigates how moral advice from AI and human advisors, its normative framing (utilitarian vs. deontological), and dilemma difficulty (easy, medium, difficult) influence participants' moral decisions and subjective experience across a range of classical moral dilemmas.

Moral dilemmas have served long before the century of AI as a cornerstone for investigating the cognitive and emotional processes underlying moral judgment. These scenarios typically pit utilitarian reasoning (maximizing outcomes) against deontological intuitions (upholding moral rules) (J. Greene & Haidt, 2002; J. D. Greene et al., 2004). While their abstract nature may limit ecological validity, classical moral dilemmas remain highly valuable for research. Their simplicity allows for the isolation of cognitive mechanisms and the testing of basic principles in moral psychology, while their controlled structure enables systematic manipulation of variables. This makes them particularly well suited for investigating how people process and respond to morally charged scenarios, especially when thoughtfully adapted to explore factors such as advice framing, dilemma complexity, and emotional gravity, as in the present study.

Recent work has begun to examine how moral decisions in these dilemmas are influenced by AI-generated advice. Individuals often follow advice from AI systems such as ChatGPT, even when the advice lacks explicit justification (Krügel et al., 2023a, 2023b, 2025). These findings raise critical questions about the persuasive power of moral AI advice, especially under conditions of cognitive strain, where users may prioritize convenience over deliberation. Moreover, research suggests that AI systems are commonly perceived to favor utilitarian decisions over deontological ones (Myers & Everett, 2025; Z. Zhang et al., 2022), a perception that may shape how advice from artificial sources is evaluated.

In parallel, the concept of transparency has emerged as a central design principle for trustworthy AI. In the moral domain, transparency involves not only revealing the source of advice but also explaining the underlying reasoning or normative assumptions. Transparent systems are thought to enhance user trust and ethical accountability (Ashoori & Weisz, 2019; Ismatullaev & Kim, 2022; Jussupow et al., 2021; You et al., 2022). This is supported by research in applied domains such as healthcare, where explainable AI has been shown to

improve trust, perceived advice quality, and even diagnostic accuracy among non-experts (Gaube et al., 2023). Crucially, explainability is particularly important in domains where AI systems have not yet earned historical trust through long-term performance, as is the case in both healthcare and moral decision-making (Cutillo et al., 2020). Yet explainability alone may not make advice more persuasive or actionable (Myers & Everett, 2025), especially when it challenges moral intuitions or increases cognitive load. It remains an open question whether the inclusion of argumentation to increase transparency actually improves the perceived credibility or effectiveness of moral advice. Another underexplored factor is the difficulty and structure of the dilemma itself. Perceived dilemma difficulty may influence both the openness to external advice and the willingness to assume responsibility. When situations are less clear or cognitively demanding, individuals may rely more on social cues, such as advice framing, and be motivated to offload responsibility, making dilemma complexity a critical factor in understanding moral advice-taking and decision-making dynamics.

Task Difficulty and High-Stakes Contexts Advice-taking does not occur in a vacuum but depends strongly on contextual factors such as task difficulty and social accountability. In high-stakes, highly public settings (e.g., televised competitions), individuals tend to overweight external recommendations, giving them more weight than their own prior judgments. This pattern likely reflects social pressure not to ignore advice and a motivation to diffuse responsibility when outcomes are consequential (Løhre & Halkjelsvik, 2024). However, this social conformity effect appears attenuated when the advisor is artificial, as such systems are not perceived as socially evaluative (Northey et al., 2022).

At the same time, not all advice weighting reflects either aversion or enthusiasm toward AI sources. Empirical findings indicate that, in many contexts, participants show neither strong algorithm aversion nor algorithm appreciation when receiving artificial advice, suggesting a more neutral, context-dependent evaluation (Gaube et al., 2021, 2023).

The influence of advice is further moderated by task difficulty. People generally underweight external input on easy tasks but increasingly rely on advice as task complexity and uncertainty rise (Gino & Moore, 2007). These findings highlight that moral advice-taking, particularly the evaluation of AI versus human input, must be interpreted in light of situational demands, cognitive load, and perceived responsibility. While some dilemmas provoke immediate emo-

tional reactions, others require complex cognitive integration or involve morally ambiguous outcomes. The degree to which users are receptive to external moral advice (including its framing and justification) may depend not only on who delivers the advice but also on the nature of the dilemma it addresses.

To address these gaps, the present study examines how individuals respond to moral advice that varies in advisor (AI vs. human) and advice direction (utilitarian vs. deontological) across classical moral dilemmas. By incorporating argumentative content (i.e., justifications or explanations for moral advice) to enhance transparency, and by varying dilemma difficulty and emotional gravity, the study aims to clarify when and how explainable advice influences both behavioral decisions and subjective experience in morally challenging contexts.

Aims and Hypotheses

By systematically varying the difficulty and structure of nine distinct scenarios, while providing explainable advice, the study aims to test the robustness of previously observed effects about the persuasive influence of artifical advice on moral decision-making. In doing so, the study seeks to replicate and extend earlier findings by e.g. Krügel et al. (2023b, 2025) while examining whether differences in dilemma difficulty, as well as advice explainability, affect participants' responsiveness to advice. In addition to behavioral responses, the study also investigates how users subjectively experience advice-based decisions, focusing on perceived ease, felt responsibility, and decision certainty. These aims address when and how AMAs influence decision-making and whether this influence differs meaningfully from that of human advisors.

Based on these objectives, we formulate the following hypotheses: First, we expect that moral advice significantly influences participants' decisions across dilemmas (*H1*). Specifically, participants should be more likely to choose a utilitarian option after receiving utilitarian advice (*H1a*), and more likely to choose a deontological option after receiving deontological advice (*H1b*), compared to a no-advice control condition. Beyond decision outcomes, we also examine how advice affects subjective aspects of the decision-making process. We expect that receiving advice (whether utilitarian or deontological, or human or artificial) will lead participants to experience the decision as easier (*H2*), perceive themselves as less personally

responsible for the outcome (*H3*), and report greater certainty in their final decision (*H4*). In addition to testing these hypotheses, the study includes exploratory analyses to examine potential factors associated with the uptake of moral advice. Specifically, we investigate whether individual differences (like wisdom and moral self-concept) are related to the likelihood of acting in line with external advice. Furthermore, we explore whether contextual factors, such as perceived dilemma difficulty, influence advice-following. As there is limited prior research on how these variables interact with moral decision support, these analyses are intended to be hypothesis-generating and serve to identify potential directions for future empirical work.

3.3 Methods

3.3.1 Participants

Participants were recruited through multiple channels, including social media platforms (e.g. Facebook, LinkedIn), online university platforms such as those of the University of Regensburg, SurveyCircle, university classes, and personal outreach. The final combined sample consisted of participants $N = 1168$ ($M_{\text{age}} = 22.06$, $SD = 7.05$; 851 women, 263 men, 10 diverse, 44 without response). The eligibility criteria required participants to be at least 18 years old and fluent in German. The sample consisted primarily of university students as we recruited a majority of the participants at the University of Regensburg. Students enrolled at the University of Regensburg were eligible to receive 0.5 participant hours (*Versuchspersonenstunden*) as compensation, in line with the university's participant credit system. No financial compensation was offered to participants recruited through other channels. All participants gave informed consent before beginning the study. Ethical approval for all studies was obtained from the Ethics Committee at the University of Regensburg Ethics Committee (approval code: 23-3326-101). Table 3.1 displays the sociodemographic composition of the sample based on gender.

3.3.2 Procedure

Upon accessing the survey, participants first provided informed consent and confirmed their agreement with the data protection regulations of the University of Regensburg. Following

Table 3.1: Sociodemographic Characteristics by Gender

Characteristics	**Total (N = 1168)**	**Male (n = 263)**	**Female (n = 851)**	**Diverse (n = 10)**
Educational attainment (%)				
No degree	6 (0.5%)	2 (0.8%)	–	4 (40.0%)
Lower secondary school[a]	1 (0.1%)	1 (0.4%)	–	–
Intermediate secondary school[a]	8 (0.7%)	3 (1.1%)	4 (0.5%)	–
Vocational high school (apprenticeship)	13 (1.1%)	3 (1.1%)	7 (0.8%)	–
Fachabitur / vocational baccalaureate	26 (2.2%)	6 (2.3%)	20 (2.4%)	–
Higher school certificate (Abitur)	967 (82.8%)	212 (80.6%)	742 (87.2%)	6 (60.0%)
University of Applied Sciences degree	20 (1.7%)	10 (3.8%)	10 (1.2%)	–
University degree	94 (8.0%)	25 (9.5%)	68 (8.0%)	–
No answer	33 (2.8%)	1 (0.4%)	–	–
Occupation (%)				
Social & Education	850 (72.8%)	170 (64.6%)	668 (78.5%)	3 (30.0%)
Technology & IT	19 (1.6%)	11 (4.2%)	6 (0.7%)	2 (20.0%)
Healthcare	38 (3.3%)	10 (3.8%)	25 (2.9%)	–
Construction & Real Estate	3 (0.3%)	1 (0.4%)	1 (0.1%)	1 (10.0%)
Goods & Services	10 (0.9%)	3 (1.1%)	5 (0.6%)	2 (20.0%)
Media	16 (1.4%)	5 (1.9%)	11 (1.3%)	–
Business & Administration	27 (2.3%)	7 (2.7%)	20 (2.4%)	–
Transport & Logistics	2 (0.2%)	1 (0.4%)	–	1 (10.0%)
Natural Sciences	56 (4.8%)	19 (7.2%)	37 (4.3%)	–
Other	112 (9.6%)	34 (12.9%)	77 (9.0%)	1 (10.0%)
No answer	35 (3.0%)	2 (0.8%)	1 (0.1%)	–

Note. [a]German school qualification levels. Gender information was missing for 44 participants and is excluded from the gender-specific columns. Percentages in the gender columns are column percentages within gender.

consent and demographic data, they received general instructions stating that they would be presented with a series of moral dilemmas and asked to make binary decisions. Participants were informed that there were no right or wrong answers and that some of the scenarios might involve morally sensitive topics. They were also provided with information about the study's duration, data privacy, and their rights as participants. To familiarize participants with the task, a sample dilemma (The Trolley Dilemma, Foot (1967)) was shown at the beginning of the survey. Participants were then randomly assigned to one of three groups: (1) AI advice, (2) human advice, or (3) control. Participants in the experimental groups received a brief, condition-specific introduction explaining the advisor providing the advice. The advisor remained consistent throughout the entire questionnaire to ensure clarity. Great care was taken to ensure that the advice introductions were comparable in tone, content, and length across conditions. Participants then proceeded to the main task. Nine dilemmas were presented one at a time in randomized order, and participants could not return to previous items once an answer had been submitted. After reading each dilemma, participants in the experimental groups received advice recommending either Option A or Option B. The content of the advice varied across dilemmas (favoring either utilitarian or deontological decisions), while the advisor (AI or human) remained fixed for each participant. Control group participants received no advice. Participants recorded their decision by selecting between two labeled options ("Option A" or "Option B") corresponding to the two action alternatives described in the dilemma. Following each decision, participants rated how easy the decision was, how responsible they felt for their decision and how certain they were about their decision. These items were presented on 7-point Likert scales (1 = not at all, 7 = very much). In addition, a behavioral variable indicating whether the participant followed the advice was computed post-hoc based on the alignment between the received advice and the selected decision. No time limit was imposed for reading, responding to advice, or making decisions. The entire survey took approximately 15–20 minutes to complete. After the main task, participants completed three standardized questionnaires in fixed order: the AI Acceptance Scale Sindermann et al. (2021), the San Diego Wisdom Scale Thomas et al. (2022), and the Moral Identity Scale Aquino and Reed (2002). These were administered at the end of the study to avoid influencing participants' decision-making during the dilemma task, as they include items related to morality, reasoning,

and attitudes towards advice. Finally, participants received a short debriefing that explained the purpose of the study, clarified that all advice was simulated rather than generated in real time, and thanked them for their participation.

3.3.3 Measures

Dependent Variables

The dependent variable was the binary choice of action (DV1). This variable describes the decision made by every participant after the confrontation with each dilemma. The second dependent variable (DV2) was advice-following, which was not measured directly but derived post-hoc from two variables in the dataset: the advice given, and the decision made. If the participant's decision matched the recommended option, the response is coded as "advice followed" (1); otherwise, it was coded as "not followed" (0). This variable reflects the proportion of advice-consistent decisions across all dilemmas.

Dilemmas

All dilemmas are adapted from established experimental research by J. D. Greene et al. (2004, 2008). To capture different levels of moral conflict, three types of dilemmas were used: three non-moral dilemmas (low-conflict, "easy" scenarios), three impersonal moral dilemmas (intermediate conflict "medium"- difficulty), and three personal moral dilemmas (high-conflict, "difficult" scenarios). A typical example for a personal moral dilemma is the widely studied trolley dilemma, which was used for explanatory purposes: "A trolley is out of control and heading toward five people standing on the tracks. You have the option to pull a lever, diverting the trolley onto another track where one person is standing. Would you pull the lever?" (German original: *"Ein Zug ist außer Kontrolle und fährt auf fünf Personen zu, die auf den Gleisen stehen. Sie können einen Hebel betätigen, der den Zug auf ein anderes Gleis umleitet, auf dem jedoch eine einzelne Person steht. Würden Sie den Hebel betätigen?"*). Selection criteria focused on the mean decision-score, aiming for a close balance between utilitarian and deontological responses. Dilemmas with current thematic social relevance were excluded (e.g., vaccination dilemmas related to the global COVID-19 pandemic and war-related dilemmas due to the Russian invasion of Ukraine), to ensure that participants'

responses were not influenced by their strong personal opinions or emotional reactions to these current events.

Table 3.2 shows the dilemmas we adapted from J. D. Greene et al. (2004, 2008) and their reported average percentage of utilitarian decisions. Some were slightly modified to create a more balanced decision.

Table 3.2: Mean Decision Percentages by Dilemma and Difficulty (J. D. Greene et al., 2004, 2008)

Difficulty	**Nonmoral**			**Impersonal Moral**			**High Conflict**		
Dilemma Title	Train or Bus	Scenic Route	Computer	Donation	Standard Fumes	Taxes	Sacrifice	Lifeboat	Safari
Mean Decision	/	/	/	24%	76%	51%	71%	22%	63%

Note. Percentages reflect the proportion of utilitarian decisions. Data adapted from Greene (2004, 2008).

Advice Stimulus

Participants in the experimental conditions (AI & human advice) received advice after reading each dilemma but before making their decision. The advice consisted of a written statement of approximately two to three sentences, recommending one of the two possible options and providing a short explanation. All advice was AI-generated, using ChatGPT (OpenAI, 2024). We provided ChatGPT with the dilemma and used prompts to develop advice that involves either utilitarian or deontological arguments. Instructions were included in the prompts to shorten the advice to 3–4 sentences and to use the subjunctive form. All advice was framed as suggestions rather than directives, thereby emphasizing the advisory rather than agentic nature of the system. While we cannot control participants' reactance, we attempted to reduce resistance through cautious formulations. All interaction with ChatGPT was in English, due to better quality of the LLM, and later translated to German. In the AI condition, the advice was framed as originating from an AI system, presented as a simulated output from ChatGPT. In the human advice condition, the advice was framed as a direct quotation from a moral expert. The content of the advice was identical across the two experimental conditions except for the framing. Participants in the control group completed all dilemmas without receiving

any advice. The assignment of advice (deontological or utilitarian) was randomized for each participant and for each dilemma. Additionally, the order of the dilemmas was randomized across participants to minimize order effects.

Additional Measures

Covariates After each dilemma, the subjective experience of the decision was assessed. We used a 7-point Likert scales (1 = not at all, 7 = very much), to assess (1) how easy they found the decision ("Wie leicht ist Ihnen die Entscheidung gefallen?"), (2) how responsible they felt for their decision ("Wie verantwortlich fühlen Sie sich für Ihre Entscheidung?"), and (3) how certain they were about their decision ("Wie sicher sind Sie in Ihrer Entscheidung?").

Established Instruments Following completion of all dilemmas, participants completed three standardized questionnaires presented in a fixed order:
First, participants completed the AI Acceptance Scale by Sindermann et al. (2021), which measures attitudes toward and trust in Artificial Intelligence. The scale consists of four items rated on a 7-point Likert scale. Higher scores reflect greater acceptance of AI systems. Second, participants completed the San Diego Wisdom Scale (SD-WISE; Sindermann et al. (2021)), which assesses individual differences in wisdom-related characteristics across five subdomains: decisiveness, emotional regulation, self-reflection, pro-social behaviour, social advising, acceptance of divergent perspectives. The SD-WISE also includes a sixth subscale called spirituality, but we excluded the related items due to a lack of relevance for our research question. The SD-WISE contains 24 items rated on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree), with higher scores indicating greater levels of wisdom. An example item is "I remain calm under pressure.". Ultimately, participants completed the Moral Identity Scale (MIS) by Aquino and Reed (2002), which measures the extent to which moral traits are central to one's self-concept. The scale comprises two subscales: Internalization, reflecting the private importance of morality to the individual, and Symbolization, reflecting the public expression of moral traits. Each subscale includes five items rated on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The measure of the MIS is twofold: First, participants are presented with a set of nine adjectives that describe a prototypically

moral person (e.g., caring, fair, honest). Subsequent, participants are asked to indicate their level of agreement with various statements about this fictitious person. An example item from the Internalization subscale is "Being someone who has these characteristics is an important part of who I am.". These questionnaires were administered at the end of the survey to avoid interfering with participants' moral decision-making processes during the main task. All materials were presented in German. The survey was administered online using SoSci Survey (Leiner, Dominik J., 2019) and was accessible via a range of devices, including personal computers, tablets, and smartphones.

3.3.4 Data Analysis

Data Preparation and Exclusion Criteria

All analyses were conducted on data collected via online questionnaires using SoSci Survey. Participants who failed to respond to individual dilemmas were retained in the analysis, as generalized linear mixed-effects models (GLMMs) are well suited to handle missing data in repeated measures designs (Hilbert et al., 2019). This modeling approach allows for the inclusion of all available data points without requiring listwise deletion, thus preserving statistical power. Dilemmas were excluded from analysis if their mean decision score across participants exceeded 0.90 or fell below 0.10, indicating that the scenario likely failed to elicit genuine moral conflict due to near-universal responses.

Experimental Design and Variable Construction

The study employed a 2 $\times$ 2 mixed factorial design with an additional no-advice control condition. Advisor (AI vs. human) was manipulated between subjects, and advice type (utilitarian vs. deontological) was manipulated within subjects; participants in the control condition received no advice throughout. Participants were randomly assigned to one of three groups and encountered different advice types across dilemmas. The advisor (AI or human) remained consistent for each participant across all dilemmas, while the type of advice varied between dilemmas. To address issues of multicollinearity we employed dummy coding, following Hilbert et al. (2019). A composite five-level condition variable was constructed to jointly reflect advisor (AI vs. human) and advice type (deontological vs. utilitarian), resulting

in five conditions: AI–deontological advice, AI–utilitarian advice, human–deontological advice, human–utilitarian advice, and a no-advice control condition.

This dummy-coded factor allowed for clear, unambiguous comparisons between all meaningful combinations of advisor and content, while avoiding multicollinearity (Hilbert et al., 2019). The primary dependent variable was the participant's binary decision in each moral dilemma. A secondary binary variable capturing advice-following was computed post-hoc. The variable was coded as 1 when the participant's decision aligned with the provided advice and 0 otherwise. Continuous secondary outcomes included perceived ease of decision, subjective responsibility, and decision certainty, each rated on a 7-point Likert scale.

Model Specification

To analyse binary outcomes (decision, advice-following), we used logistic mixed-effects models. For continuous secondary variables (ease, responsibility, certainty), we used linear mixed-effects models. All models included fixed effects for the five-level condition variable and random intercepts for participants and dilemmas to account for baseline variation across individuals and decision scenarios. The general structure of the GLMM was as follows:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1(\text{Condition}_{ij}) + u_{0i} + v_{0j} \tag{3.1}$$

Where:

- π_{ij} is the probability of a particular decision for participant i on dilemma j
- β_0 is the fixed intercept
- β_1 represents the fixed effect of the five-level condition variable
- u_{0i} is the random intercept for participant i
- v_{0j} is the random intercept for dilemma j

This model accounts for both the fixed effect of experimental condition and the random structure imposed by repeated measures across dilemmas and participants. For linear outcomes, the structure was analogous but implemented using lmer(). Mixed-effects modeling was selected over traditional ANOVA approaches due to its superior handling of hierarchical data and

missing responses (Al-Dajani & Czyz, 2022; Hilbert et al., 2019). All models were checked for convergence and multicollinearity. Model performance was evaluated using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Although formal model comparison was not the primary aim of the study, these indices were used as relative information criteria to assess model adequacy and to guard against overfitting.

Software and Implementation

All statistical analyses were conducted in R using Posit Software (2025). Data were imported from Excel files and managed using the `readxl` and `openxlsx` packages, and pre-processed with functions from the `tidyverse` ecosystem, including `dplyr` and `tidyr`, for data cleaning, recoding, and the construction of derived variables. Generalized linear mixed-effects models and linear mixed-effects models were estimated with the `lme4` package, and model-based visualizations were created using `ggplot2`. Estimated marginal means and predicted probabilities for the graphical presentation and interpretation of interaction effects were obtained using the `ggemmeans` package.

Statistical Thresholds and Reporting

All significance tests were evaluated using a conventional alpha level of 0.05. For binary outcomes, results are reported as odds ratios (OR) along with 95% confidence intervals. No correction for multiple comparisons was applied, as all analyses were based on preregistered hypotheses. To enhance the interpretability of logistic mixed-effects model results, odds ratios were computed by exponentiating the unstandardized regression coefficients (i.e., $\text{OR} = e^{\beta}$). Odds ratios offer an intuitive measure of effect size, indicating how the odds of the outcome event (advice-following) change with a one-unit increase in the predictor variable. This method is particularly suitable for binary dependent variables analyzed using logistic regression (J. Hilbe, 2009; J. M. Hilbe, 2009). All reported odds ratios are accompanied by their corresponding p-values and are interpreted relative to the reference category for categorical predictors. This facilitates meaningful comparisons across experimental groups and psychological predictors, providing a clearer understanding of their practical relevance for decision-making behavior.

Post-hoc Sensitivity Analyses

Post-hoc power analysis was conducted for Study 1 using G*Power 3.1 Faul et al. (2007) to evaluate whether the available sample size was sufficient to detect meaningful effects in a logistic regression model. Assuming a medium effect size (odds ratio = 2.5), a one-tailed alpha level of 0.05, and approximately balanced group sizes ($\pi = 0.5$), the statistical power for Study 1 ($N = 1168$) was estimated at 1.00. This indicates excellent sensitivity to detect effects of the hypothesized magnitude, providing strong support for the adequacy of the sample size.

3.4 Results

3.4.1 Summary

The present study provides strong evidence, that the participants' moral decisions were influenced by the moral advice they received. Specifically, deontological advice, whether attributed to an AI or a human expert, significantly increased the likelihood of participants choosing deontological actions in response to moral dilemmas. In contrast, utilitarian advice did not produce a significant shift in decision-making. Advice-following behavior was relatively consistent across both advisors, indicating no major differences in persuasive impact between AI and human advisors. Individual differences also played a role: participants with higher levels of wisdom were more likely to follow advice overall, while moral identity internalization and wisdom were found to influence perceived ease, certainty, and responsibility in nuanced ways. Exploratory analyses further showed that the presence of advice affected perceived responsibility in easier dilemmas but had limited impact on other subjective decision metrics. Overall, the results suggest a persuasive asymmetry favoring deontological over utilitarian advice, regardless of the advisor.

3.4.2 Sample Characteristics

A total of N = 1,168 (M_{age} = 28.54, SD = 11.13) participants took part in the study, contributing a total of 7,444 individual dilemma responses. Sample characteristics are reported in Table 3.1. The gender distribution was: 851 female (72.9%), 263 male (22.5%), 10 identifying as diverse (0.9%), and 44 participants (3.8%) who did not report their gender.

3.4.3 Mean Decision and Advice-Following

Participants' decisions varied systematically across the eight moral dilemmas. Dilemma "computer" was excluded from all analyses due to a highly skewed response distribution (with >90% selecting the same option, as displayed in table 3.3), indicating it did not evoke a meaningful moral conflict. In the high-conflict personal dilemmas, deontological response rates were moderate to low: dilemma "sacrifice" ($M = 0.667$, $SD = 0.472$), dilemma "lifeboat" ($M = 0.354$, $SD = 0.571$), and dilemma "safari" ($M = 0.288$, $SD = 0.453$). The impersonal dilemmas showed broader variability, with dilemma "donation" ($M = 0.304$, $SD = 0.460$), dilemma "standard fumes" ($M = 0.307$, $SD = 0.462$), and dilemma "taxes" ($M = 0.578$, $SD = 0.584$). In the non-moral dilemmas, participants predominantly chose the deontological option in dilemma "train/ bus" ($M = 0.717$, $SD = 0.455$) and dilemma "scienic route" ($M = 0.709$, $SD = 0.450$). Regarding mean advice-following, participants followed advice in more than half of the trials, with highly consistent adherence across conditions. In the AI condition, the overall advice-following was $M = 0.548$ ($SD = 0.498$), while participants in the human advice condition followed the advice with a mean rate of $M = 0.541$ ($SD = 0.498$). When broken down by dilemma, advice-following in the AI condition was highest in dilemma "sacrifice" ($M = 0.590$, $SD = 0.492$) and lowest in dilemma "safari" ($M = 0.519$, $SD = 0.500$). The other dilemmas showed moderate adherence: dilemma "train/ bus" ($M = 0.562$, $SD = 0.497$), dilemma "scienic route" ($M = 0.542$, $SD = 0.497$), dilemma "donation" ($M = 0.549$, $SD = 0.498$), dilemma "standard fumes" ($M = 0.546$, $SD = 0.498$), dilemma "taxes" ($M = 0.567$, $SD = 0.497$), and dilemma "lifeboat" ($M = 0.547$, $SD = 0.498$). In the human advice condition, the highest advice-following occurred in dilemma "safari" ($M = 0.571$, $SD = 0.496$), followed closely by dilemma "sacrifice" ($M = 0.595$, $SD = 0.490$) and dilemma "standard fumes" ($M = 0.570$, $SD = 0.496$). The remaining dilemmas yielded the following advice-following rates: dilemma "train/ bus" ($M = 0.579$, $SD = 0.494$), dilemma "scienic route" ($M = 0.530$, $SD = 0.500$), dilemma "donation" ($M = 0.530$, $SD = 0.500$), dilemma "taxes" ($M = 0.514$, $SD = 0.500$), and dilemma "lifeboat" ($M = 0.545$, $SD = 0.499$). Overall, these results indicate that both the difficulty level and moral framing of the dilemmas influenced decision tendencies, while advice-following remained relatively stable across both advisors. A summary of aggregated results is shown in Table 3.3.

Table 3.3: Decision Rates and Advice-Following by Dilemma Type and Group

Group	**Nonmoral**			**Impersonal Moral**			**Personal Moral**		
	Train/Bus	Computer	Scenic Route	Donation	Standard Fumes	Taxes	Sacrifice	Lifeboat	Safari
Mean decision (literature)	/	/	/	0.37	0.24	0.76	0.49	0.29	0.78
All	0.72	0.039	0.71	0.30	0.31	0.59	0.67	0.36	0.29
	(0.45)	(0.20)	(0.46)	(0.46)	(0.46)	(0.49)	(0.47)	(0.48)	(0.45)
AI Advice	0.72	0.04	0.72	0.31	0.33	0.56	0.66	0.37	0.30
	(0.45)	(0.19)	(0.45)	(0.46)	(0.47)	(0.50)	(0.48)	(0.48)	(0.46)
Human Advice	0.72	0.05	0.72	0.32	0.30	0.62	0.69	0.38	0.29
	(0.45)	(0.21)	(0.45)	(0.47)	(0.46)	(0.49)	(0.46)	(0.49)	(0.45)
Control Group	0.71	0.02	0.62	0.22	0.24	0.61	0.61	0.26	0.24
	(0.46)	(0.15)	(0.49)	(0.42)	(0.43)	(0.49)	(0.49)	(0.44)	(0.43)
AI Advice-Following	0.56	0.52	0.54	0.55	0.55	0.56	0.59	0.55	0.52
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)	(0.50)
Human Advice-Following	0.58	0.51	0.53	0.53	0.57	0.51	0.52	0.55	0.57
	(0.49)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)

Note. Values represent means with standard deviations in parentheses.

3.4.4 Main Analysis

Empty Model

To assess the amount of variance attributable to differences between participants and dilemmas, an empty multilevel logistic regression model (i.e., intercept-only model) was estimated. This model included random intercepts for both participants and dilemmas but no fixed predictors. The model showed good convergence and no warnings. The variance component for participants was $\sigma^2 = 0.361$ ($SD = 0.601$), and for dilemmas $\sigma^2 = 0.686$ ($SD = 0.828$), indicating that a substantial proportion of variance in decisions was attributable to both participant-level and dilemma-level differences. These results justify the use of generalized linear mixed-effects models (GLMMs) in subsequent analyses. The fixed intercept was not significantly different from zero ($\beta = -0.03$, $SE = 0.29$, $z = -0.11$, $p = .914$), suggesting that the baseline log-odds of making a utilitarian decision (across all dilemmas and participants) were approximately balanced. The Akaike Information Criterion (AIC = 9260.7) and Bayesian Information Criterion (BIC = 9281.4) from this null model were used as a baseline for evaluating model fit in subsequent hypothesis-driven models.

The Influence of Advice and Advisor on Decision-Making

To test Hypothesis *1a* and*1b*, a generalized linear mixed-effects model (GLMM) was conducted using a binomial distribution and logit link function to predict participants' decisions in moral dilemmas. The binary outcome variable (decision) was coded such that the reference level was utilitarian, meaning the model estimates the log-odds of choosing a deontological response (Option B) over a utilitarian one (Option A). Fixed effects included the experimental condition (advisor), dummy-coded to compare each advice condition to the control group (no advice). Random intercepts were included for participants and dilemmas to account for within-subject and within-stimulus variance.

The model showed acceptable convergence and no singularity or boundary warnings. The fixed intercept ($\beta = -0.31$, $SE = 0.33$, $z = -0.96$, $p = .336$) represents the baseline log-odds of choosing a deontological response in the control group and was not significantly different from zero, indicating roughly even odds of utilitarian and deontological decisions in the

absence of advice.

Participants who received deontological advice from either advisor were significantly more likely to choose the deontological option compared to the control group. Specifically, deontological AI advice significantly increased the likelihood of a deontological decision ($\beta = 0.56$, $SE = 0.12$, $z = 4.55$, $p < .001$, $OR = 1.75$), as did deontological advice from a human advisor ($\beta = 0.58$, $SE = 0.12$, $z = 4.65$, $p < .001$, $OR = 1.78$). These results indicate that the odds of making a deontological decision were approximately 75–78% higher when participants were advised deontological, regardless of whether the advice came from a human or AI. This effect is visually illustrated in Figure 3.1, which presents the predicted probabilities of a deontological decision across advice conditions, and in Figure 3.2, which shows the corresponding odds ratios with confidence intervals.

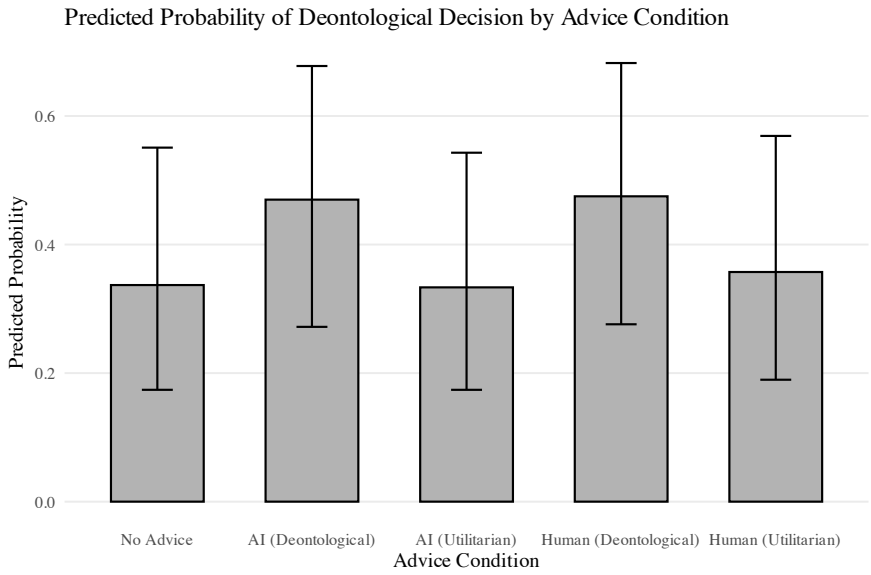


Figure 3.1: Predicted probability of egoistic decisions across advice conditions. Error bars represent 95% confidence intervals.

In contrast, utilitarian advice from either advisor did not significantly affect participants' decisions relative to the control condition. This was also stable across both advisors (AI: $\beta = -0.02$, $SE = 0.10$, $z = -0.19$, $p = .853$, $OR = 0.98$; human: $\beta = 0.08$, $SE = 0.13$, $z = 0.66$, $p = .509$, $OR = 1.08$), both yielding non-significant effects. The odds of choosing the utilitarian option did not differ significantly from the control group. The odds ratios close to 1 (AI: $OR = 0.98$; human: $OR = 1.09$) indicate a negligible change in the likelihood

of a utilitarian decision, suggesting that utilitarian advice was largely ineffective in shifting participants' decisions.

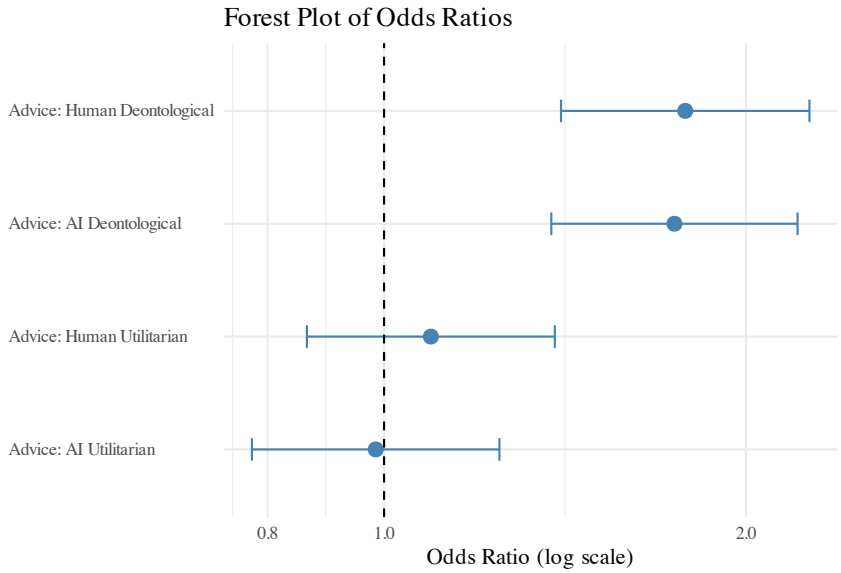


Figure 3.2: The effects of advice and advisor on the likelihood of a deontological decision. A value above 1 indicates increased odds relative to the control condition.

The model was based on 7,444 observations from 1,036 participants across eight dilemmas. Random intercept variances were $\sigma^2 = 0.36$ for participants and $\sigma^2 = 0.72$ for dilemmas, confirming that both individual and stimulus-level variability influenced decision behavior and justifying the use of multilevel modeling. Model fit statistics were acceptable (AIC = 9167.0, BIC = 9215.4). The model explained a small to moderate proportion of variance in participants' decisions, with a marginal $R^2 = .02$ and a conditional $R^2 = .26$. Full model results are presented in Table 3.4.

Table 3.4: Logistic Mixed-Effects Model for Decision Prediction

Predictor	β	**SE**	**z**	**p**	**OR [95% CI]**
Intercept (Control group)	−0.310	0.330	−0.96	.336	0.74 [0.38, 1.38]
AI Advice – Deontological	0.560	0.120	4.55	**<.001**	1.75 [1.38, 2.22]
AI Advice – Utilitarian	−0.020	0.100	−0.19	.853	0.98 [0.80, 1.21]
Human Advice – Deontological	0.580	0.120	4.65	**<.001**	1.78 [1.40, 2.27]
Human Advice – Utilitarian	0.080	0.130	0.66	.509	1.09 [0.85, 1.37]

Note. Reference category: Control (no advice). β = unstandardized log-odds coefficient; SE = standard error; OR = odds ratio; CI = confidence interval. Bolded p values indicate statistical significance at $p < .05$.

Why We Follow Advice

A logistic mixed-effects regression was conducted to examine the influence of moral identity, AI acceptance, advice type, and advisor on advice-following (DV2). The dependent variable was whether participants followed the advice (1) or not (0). Fixed effects included internalization (scaled), AI acceptance, advice type (utilitarian vs. deontological), advisor (AI vs. human), difficulty, and wisdom, as well as interaction terms between group and both internalization and AI acceptance. Random intercepts were included for participants and dilemmas. The analysis revealed that wisdom significantly predicted advice-following, $\beta = 0.15$, $SE = 0.07$, $z = 2.06$, $p = .039$. No other predictors, including internalization, AI acceptance, advice type, or advisor, reached significance. This suggests that individuals with higher levels of wisdom are generally more inclined to act in accordance with external moral advice, regardless of its advisor or framing. We found no significant main effects of dilemma difficulty (impersonal: $p = .533$, $OR = 0.95$; personal: $p = .993$, $OR = 1.00$), nor a significant difference between the human and AI advice groups ($p = .354$, $OR = 0.95$). Furthermore, none of the interaction terms reached significance (e.g., human × impersonal: $p = .590$, $OR = 1.06$), indicating that advice-following remains stable across dilemma types or between advisors. The model accounted for very little variance in advice-following ($R^2_m = .001$, $R^2_c = .003$). All results are displayed in Table 3.5.

Table 3.5: Logistic Mixed-Effects Model Predicting Advice-Following

Predictor	β	SE	z	p	95% CI (Wald)
Intercept	−0.33	0.31	−1.05	.294	[−0.94, 0.29]
Advisor: Human	−0.05	0.05	−0.93	.354	[−0.16, 0.06]
Advice Type: Utilitarian	0.00	0.05	0.06	.949	[−0.10, 0.11]
Internalization	0.02	0.05	0.33	.742	[−0.08, 0.11]
AI Acceptance	−0.01	0.03	−0.19	.847	[−0.06, 0.05]
Wisdom	0.15	0.07	2.06	**.039**	[0.01, 0.29]
Difficulty: Impersonal	−0.04	0.07	−0.53	.600	[−0.17, 0.10]
Difficulty: Personal	−0.04	0.07	−0.61	.544	[−0.18, 0.09]

Note. Reference categories: Advice Type = Deontological; Advidsor = AI; Difficulty = Nonmoral. β = unstandardized log-odds coefficient; SE = standard error; CI = confidence interval. Bold p values indicate significance at $p < .05$.

3.4.5 Exploratory Analysis of Covariates

Ease

To test Hypothesis *2*, we used a mixed-effects model predicting perceived ease included the fixed effects of advice presence, wisdom, and internalization, with random intercepts for participant and dilemma. In the full sample, higher wisdom was significantly associated with lower perceived ease of decision-making ($\beta = -0.319$, OR = 0.73, $p < .001$), while internalization significantly increased ease ($\beta = 0.205$, OR = 1.23, $p < .001$). Advice presence did not predict ease across all dilemmas. When analyzed by difficulty, easy dilemmas showed a significant negative effect of wisdom on ease ($\beta = -0.446$, OR = 0.64, $p < .001$), whereas internalization did not predict ease. In medium-difficulty dilemmas, wisdom again reduced ease ($\beta = -0.360$, OR = 0.70, $p < .001$), and internalization significantly increased ease ($\beta = 0.170$, OR = 1.19, $p = .008$). In high-difficulty dilemmas, advice presence significantly reduced ease ($\beta = -0.326$, OR = 0.72, $p = .046$), and internalization had the strongest positive effect ($\beta = 0.349$, OR = 1.42, $p < .001$), while wisdom did not show a significant effect. The results are visualized in Figure 3.3

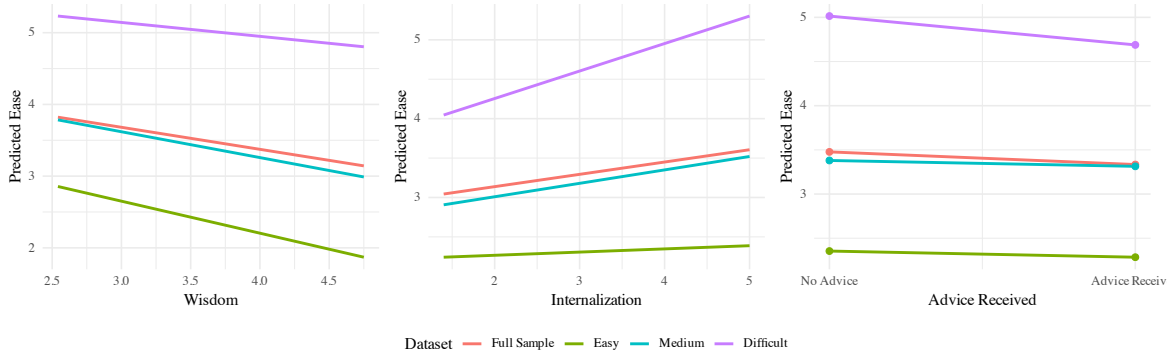


Figure 3.3: Ease of the decision. Predicted ease as a function of wisdom, internalization, and advice received across all difficulties.

Certainty

To assess Hypothesis *3*, we assessed a similar mixed-effect model to the previous assessment of subjective ease. In the full sample, wisdom was a strong negative predictor of certainty ($\beta = -0.518$, OR $= 0.60$, $p < .001$), while internalization showed a marginally significant positive effect ($\beta = 0.105$, OR $= 1.11$, $p = .054$). Advice presence was not a significant predictor. When broken down by difficulty, the effect of wisdom remained negative and significant in all conditions: easy ($\beta = -0.508$, $p < .001$), medium ($\beta = -0.524$, $p < .001$), and difficult ($\beta = -0.518$, $p < .001$). Internalization positively predicted certainty in medium difficult dilemmas ($\beta = 0.118$, $p = .048$), while remaining nonsignificant in others. Advice presence did not predict certainty in any subgroup.

As illustrated in Figure 3.4, predicted certainty increases with both wisdom and internalization, with modest differences across advice conditions.

Responsibility

Perceived responsibility as mentioned in Hypothesis 4 was examined using a third mixed effect model. Across the full sample, wisdom significantly reduced responsibility ($\beta = -0.338$, OR $= 0.71$, $p < .001$), and internalization also had a significant negative effect ($\beta = -0.188$, OR $= 0.83$, $p < .001$). Advice presence showed a marginally positive effect ($\beta = 0.177$, OR $= 1.19$, $p = .068$). In easy dilemmas, the presence of advice significantly decreased subjective responsibility ($\beta = 0.377$, OR $= 1.46$, $p = .017$) for the decision, while wisdom

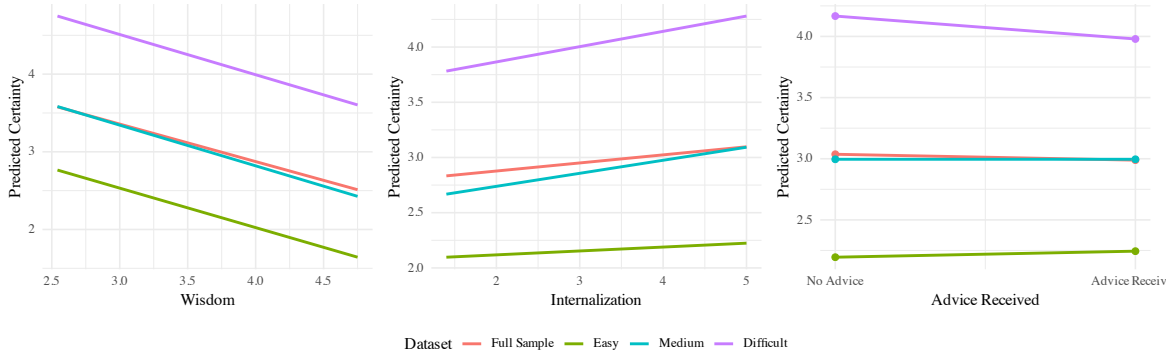


Figure 3.4: Decision certainty. Predicted certainty as a function of wisdom, internalization, and advice received across all datasets.

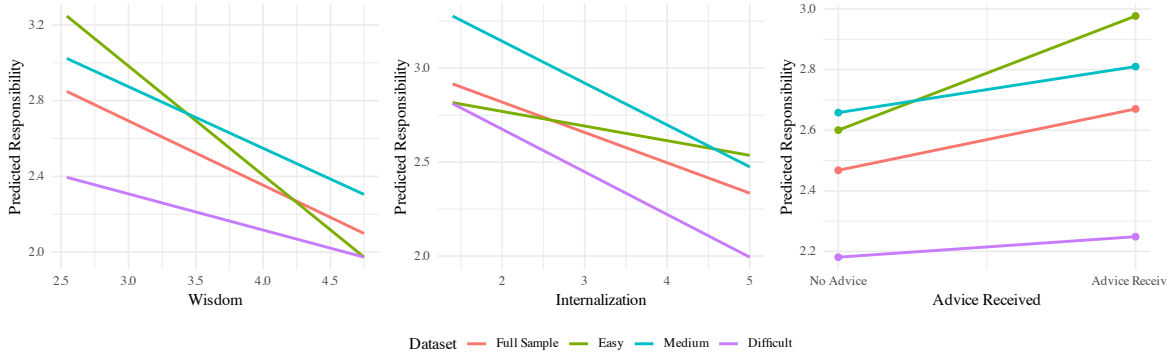


Figure 3.5: Predicted responsibility as a function of wisdom, internalization, and advice received across all datasets.

significantly improves responsibility. In medium-difficulty dilemmas, the presence of advice did not reach significance, but wisdom and internalization retained a strong negative association ($\beta = -0.325$, $p < .001$), resulting in greater perceived responsibility. In high-difficulty dilemmas, internalization again significantly increased responsibility ($\beta = -0.227$, $p < .001$), with marginal effects for wisdom and non-significant results for the presence of advice. Figure 3.5 displays predicted responsibility patterns across the same predictors, highlighting stronger effects in easier dilemmas.

3.5 Discussion

3.5.1 Summary of Key Findings

The present study examined how the advisor and content of moral advice influence decision-making in classical moral dilemmas. The findings indicate that deontological advice significantly increased deontological decisions, regardless of whether the advice came from an AI or a human. In contrast, utilitarian advice did not significantly influence participants' decisions. This suggests an asymmetry in the persuasiveness of moral framings. Moreover, wisdom emerged as a significant predictor for advice-following, suggesting that individuals high in wisdom are more likely to engage constructively with external advice. Advice-following behavior was consistent across dilemma types and advisors, but depended on the wisdom score of the individual. Additionally, advice increased decision ease in difficult dilemmas and reduced perceived responsibility in easy dilemmas, while decision certainty was driven primarily by individual wisdom rather than external advice.

3.5.2 Interpretation of the Results

Persuasive Asymmetry, Advisor Irrelevance and Utilitarian Ineffectiveness

Our findings provide important insights into the cognitive processing of moral advice. The results from the present study are in line with previous research, regarding the advisor irrelevance (Ikeda, 2024; Köbis et al., 2021; Krügel et al., 2022, 2023a, 2025; Leib et al., 2021, 2023). Surprisingly, utilitarian advice failed to significantly influence moral decisions, in contrast to deontological advice, which reliably increased deontological responses. Thus, we could not replicate the significant influence of utilitarian advice on moral decision-making found in previous literature. This influential discrepancy of utilitarian and deontological advice aligns our results more closely with Myers and Everett (2025), who showed that utilitarian advice is less trusted, even though users expect AI to favor such reasoning, and is further consistent with evidence that rule-based (deontological) advice can be relatively more effective than identity- or role-based moral appeals in Human–Robot ethical interactions (B. Kim et al., 2021). This aligns with the broader expectation that AI advisors tend to favor utilitarian decisions more readily than human decision-makers (Y. Zhang et al., 2023; Z. Zhang et al.,

2022). Yet, as both our results and prior work suggest, such expectations do not necessarily translate into greater persuasive power, particularly when utilitarian reasoning conflicts with intuitive moral norms.

A possible explanation for this discrepancy may lie in the effect of additional argumentation itself. While intended to increase transparency or rational appeal, elaborated reasoning may unintentionally draw attention to the discomfort associated with instrumental harm, thereby triggering moral resistance. As a result, persuasion may be reduced in cases where the advice includes justification or more detailled argumentation, compared to studies that omit such elaboration, as demonstrated by e.g., Krügel et al. (2022, 2023c). This asymmetry also aligns with dual-process models of moral reasoning (J. D. Greene et al., 2004), which propose that deontological judgments are emotionally grounded and processed intuitively, making them more amenable to persuasive cues, especially under cognitive strain.

These findings raise the question of why artificial advice can exert such influence on moral decision-making. One potential explanation lies in the nature of cognitive processing under conditions of limited motivation or capacity. According to the Elaboration Likelihood Model (ELM), persuasive cues such as moral advice are often processed via either a central (effortful) or peripheral (cue-based) route, depending on motivation and cognitive capacity (Petty & Cacioppo, 1986). Given the online survey format and the complexity of high-stakes dilemmas used in this study, it is likely that many participants relied on this lower-effort processing mode.

This interpretation is supported by findings from Köbis et al. (2021), who argue that AI systems are increasingly adopted in roles that enable task delegation, including ethically sensitive decisions. This tendency suggests not only a willingness to outsource moral judgment but also a potential self-limitation of cognitive engagement in such domains, as users may come to rely on external systems to reduce the mental burden of ethical deliberation.

Building on this perspective, we propose that externally provided advice, regardless of whether it originates from a human or an AI, often functions as a heuristic or mental shortcut that facilitates cognitive ease in morally demanding situations. Rather than serving as input for thorough deliberation, such advice may be adopted as a mental shortcut, especially under conditions of cognitive strain or motivational fatigue.

This argument aligns with the effort-reduction framework proposed by Shah and Oppenheimer (2008), which describes heuristics as tools for minimizing mental workload. In this context, advice use appears to reduce effort by narrowing the range of considered options, simplifying the integration of information, lowering demands on memory retrieval, and limiting attention to only the most salient cues. These features make moral advice a cognitively efficient strategy for navigating complex decisions. Following this logic, unelaborated advice, particularly utilitarian advice, may prove more persuasive because it reduces the cognitive cost of processing. In contrast, advice that includes justification may trigger deeper scrutiny, especially when it draws attention to morally uncomfortable trade-offs such as instrumental harm. This could explain the reduced effectiveness of utilitarian advice when accompanied by argumentation. While transparency is generally considered beneficial for user trust and system credibility (Ashoori & Weisz, 2019; Ismatullaev & Kim, 2022), the findings suggest that under conditions of cognitive strain, such as high-conflict moral dilemmas, argumentation may reduce the persuasive impact of AI-generated advice, particularly when it contradicts intuitive moral norms. Crucially, in such cases, the added reasoning does not sufficiently reduce cognitive complexity to enhance persuasiveness; instead, it may amplify discomfort and cognitive load. Moreover, by encouraging acceptance of counterintuitive outcomes, such as endorsing instrumental harm, elaborated utilitarian argumentation may clash with entrenched societal moral foundations, thereby diminishing its influence.

Wisdom as a Significant Predictor in Advice-Following

Wisdom is a significant predictor for advice-following. Consequently, participants who score higher on the wisdom scale tend to follow the advice more often. While this may seem counterintuitive at first, it aligns with the idea that wise individuals are typically more open to diverse perspectives, emotionally balanced in the face of complex decisions, and motivated by a concern for others (Thomas et al., 2019). These qualities likely enable them to engage with external moral advice in a reflective and constructive manner, rather than dismissing it outright. Returning to the perspective of the Elaboration Likelihood Model (Petty & Cacioppo, 1986), wisdom may reflect a greater capacity or ability to engage in central route processing, allowing for deeper evaluation of persuasive input when motivation is present.

Rather than accepting advice heuristically, these individuals may be more likely to reflect on its normative relevance and integrate it into their moral reasoning.

Subjective Experience: Ease, Responsibility, and Certainty

Ease Our exploratory analyses revealed important insights into how individual traits and situational factors shape the subjective experience of moral decision-making. Regarding decision ease, wisdom was associated with greater perceived ease in low- and medium-difficulty dilemmas, but not in high-stakes scenarios. This suggests that wisdom facilitates decision processing when cognitive load is moderate, potentially due to enhanced emotional regulation and tolerance for complexity. However, in high-conflict situations, even wise individuals may reach the limits of cognitive integration, indicating that the demands of severe moral trade-offs override the buffering effect of wisdom. Notably, advice significantly increased perceived ease in difficult dilemmas. This implies that external guidance reduces cognitive strain particularly when decisions are most challenging, likely by simplifying trade-offs or reinforcing normative direction. In contrast, advice did not significantly affect ease in easier dilemmas, where internal resources may suffice for confident reasoning. In low and medium difficulty dilemmas, wisdom appears to provide sufficient internal resources such as emotional regulation, nuanced perspective-taking, and tolerance for ambiguity, enabling relatively confident and efficient decisions. These individuals likely draw on a repertoire of reflective strategies and moral insights, making external input less necessary in less difficult situations. However, in high-conflict scenarios, the complexity and emotional weight of moral trade-offs likely exceed even those internal capacities. Here, wisdom's buffering effect likely plateaus, and advice steps in as a compensatory aid. This shift suggests a form of adaptive calibration: when internal faculties reach their cognitive or emotional limits, individuals become more receptive to structured external advice, especially if it reduces ambiguity or provides a clear normative anchor. These findings suggest that advice and wisdom play complementary roles in decision-making. While wisdom may support independent judgment under favorable conditions, advice appears to be relied upon more when internal resources are limited. Consistent with previous results, individuals with higher wisdom scores tend to use advice more frequently than those with lower scores.

Certainty As for decision certainty, only wisdom emerged as a reliable predictor. Participants scoring higher in wisdom reported greater confidence in their decisions, independent of advice or moral identity. This may reflect their broader perspective-taking and reduced susceptibility to moral paralysis in ambiguous situations. In contrast, neither receiving advice nor moral identity internalization significantly influenced certainty.

Responsibility Across all dilemmas, higher scores in moral identity internalization were significantly associated with higher perceived responsibility. This pattern was stable across difficulty levels, suggesting that individuals with a strongly internalized moral self tend to consistently attribute moral responsibility to themselves when making ethical decisions. Wisdom was also associated with increased perceived responsibility, but only in low- and medium-difficulty dilemmas. In high-difficulty scenarios, this relationship weakened and no longer reached statistical significance, likely because such scenarios involved unavoidable moral harm to one or more individuals, elevating responsibility judgments across the board. In other words, while individuals low in wisdom reported less responsibility in less demanding contexts, the inherent gravity of high-conflict dilemmas elicited uniformly high levels of perceived responsibility, reducing variability and thereby masking individual differences. Receiving advice, by contrast, significantly reduced perceived responsibility only in easy dilemmas. In medium and high-difficulty dilemmas, the effect of advice on responsibility ratings was not statistically significant. This suggests that advice may buffer perceived responsibility when decisions are relatively manageable, but its mitigating influence diminishes as the moral and emotional weight of the dilemma and the inevitability of harm increases.

3.5.3 Practical and Ethical Considerations and Implications

These findings carry important implications for the deployment of AI in moral decision-making contexts for theory development, applied system design, and ethical governance. AI systems can exert persuasive influence on moral decision-making, particularly when they mirror widely accepted social norms. While this demonstrates the potential of AMAs as moral decision aids, it also raises ethical concerns. If persuasive framing can determines behavior, systems must be carefully designed to avoid unreflective moral outsourcing or behavioral dependence.

Transparency, user autonomy, and ethical literacy remain key design principles. The findings of the present study offer multifaceted insights into how moral advice, particularly when generated or mediated by artificial systems, interacts with individual cognition, normative framing, and responsibility attribution.

Theoretical Implications

This study contributes to the growing literature on moral persuasion and AI-supported decision-making by reinforcing the importance of normative content over advisor credibility. Despite widespread discourse around AI skepticism (Bigman & Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2015; Hou & Jung, 2021; Jussupow et al., 2020; Mahmud et al., 2022), our findings suggest that participants followed deontological advice irrespective of whether it originated from a human or artificial advisor. This lends empirical support to emerging perspectives arguing that moral framing, but not the messenger, determines persuasive power in ethical contexts (Krügel et al., 2025; Myers & Everett, 2025).

Furthermore, the observed asymmetry in advice effectiveness aligns with dual-process models of moral cognition (J. D. Greene et al., 2004), which posit that deontological intuitions are more affectively grounded and readily accessible under cognitive strain. Utilitarian reasoning, in contrast, often requires effortful, abstract deliberation, making it more susceptible to resistance, particularly when justification draws attention to morally aversive outcomes such as instrumental harm. The reduction in utilitarian persuasion upon the inclusion of argumentation further illustrates the cognitive cost of elaborative moral reasoning and its diminished impact under pressure. These findings offer a meaningful extension of the Elaboration Likelihood Model (Petty & Cacioppo, 1986), illustrating that the model's core distinction between central and peripheral processing continues to apply in the domain of AI-mediated moral reasoning. Individual traits such as wisdom appear to modulate the likelihood of engaging in central route processing: participants high in wisdom were more inclined to reflectively engage with advice, integrating it with their broader moral understanding rather than accepting it as a heuristic shortcut. Yet, in high-difficulty dilemmas, even these individuals seemed to approach cognitive limits, suggesting that the threshold for effortful processing is shaped not only by personal capacity but also by situational demands.

Similarly, the role of advice as a cognitive simplification mechanism aligns with the effort-reduction principles proposed by Shah and Oppenheimer (2008). The way participants used advice, to narrow options, simplify integration, and focus on salient cues, illustrates how established theories of heuristic decision-making remain relevant in the context of AMAs. This suggests that core psychological models of persuasion and judgment retain their explanatory value even as new technological intermediaries, like AI systems, enter morally sensitive domains.

Practical Implications

These theoretical insights yield concrete recommendations for the design of Artificial Moral Advisors (AMAs) and decision-support systems. Moral advice appears most persuasive when aligned with intuitive social norms, particularly in contexts involving time pressure or emotional strain (Bleher & Braun, 2022; Constantinescu et al., 2022; Giubilini & Savulescu, 2018). In such situations, the framing of moral advice proves more influential than the identity of the advisor; specifically, appeals to deontological principles tend to elicit greater user compliance, regardless of whether the advisor is human or artificial.

The limited persuasive impact of utilitarian advice when accompanied by justification further suggests the need for adaptive transparency. Although explainability remains crucial for establishing system credibility (Lima et al., 2021), excessive elaboration when it highlights morally aversive trade-offs, can provoke discomfort or cognitive disengagement. This underscores the importance of context-sensitive interfaces: detailed reasoning should be available for users with the motivation and capacity for reflective processing (e.g., experts or individuals high in wisdom), while concise, emotionally attuned summaries may be more effective for users operating under cognitive or emotional strain.

Emerging evidence on individual differences further points to the value of user-adaptive moral guidance. Tailoring moral support to users' cognitive styles and reflective capacities can enhance both uptake and ethical engagement. For instance, those high in wisdom may benefit from prompts that encourage deeper reflection or offer access to underlying normative justifications, whereas users facing overload or stress may respond more effectively to simplified normative anchors.

A promising direction for personalization involves designing AMAs as quasi-relativistic systems, which are capable of aligning their recommendations with the user's own ethical commitments. Through structured interaction, the system can elicit moral priorities, retain them, and subsequently generate context-sensitive advice grounded in the individual's value framework (Giubilini & Savulescu, 2018). This approach allows AMAs to provide advice that resonates with users' moral intuitions, even in the absence of an explicit normative theory. Yet such personalization is not without risk. By reflecting and reinforcing preexisting beliefs, these systems may entrench unexamined biases or inconsistencies, potentially generating unbalanced or ethically problematic advice.

This risk reveals a more fundamental design challenge: AMAs should not function merely as instruments of moral alignment, but as facilitators of moral development. The goal should not be to echo the user's current worldview, but to cultivate reflective engagement and ethical growth. One productive strategy involves presenting users with counterintuitive or dissonant perspectives that challenge their assumptions. When delivered thoughtfully, such interventions can initiate a process of reflective equilibrium, encouraging individuals to revise or refine their moral beliefs. In this capacity, AMAs become not only advisors but co-participants in moral self-examination, promoting a more resilient, critically engaged ethical mindset.

Ethical Implications

The persuasive capacity of AI-generated moral advice raises significant ethical concerns, foremost among them the risk of moral outsourcing. When users process AI advice heuristically, therefore accepting advice without reflective engagement, they may gradually disengage from their own ethical reasoning. This can give rise to a diffusion of responsibility (Bleher & Braun, 2022; Waytz et al., 2014), a phenomenon long recognized in social psychology. Analogous to the bystander effect, where the presence of others diminishes individual accountability in morally salient situations, the availability of a seemingly objective and rational moral advisor may lead users to defer responsibility to the system itself. The result is a reduced perceived obligation to deliberate personally about moral decisions. Our empirical findings reinforce this concern, showing that moral advice in easy or ambiguous contexts can significantly diminish users' perceived responsibility (Formosa & Ryan, 2021; Giroux et al., 2022; Köbis et al.,

2021; Leib et al., 2023). When individuals repeatedly rely on AI for moral advice, especially in routine decision-making, their sense of moral authorship may erode. Over time, such reliance risks weakening the internalization of moral norms and diminishing motivation for autonomous ethical reasoning. In this light, the AMA does not merely function as a support system; it may evolve into a proxy moral agent, subtly shifting moral accountability away from the individual and thereby undermining ethical agency. Compounding this risk is the asymmetrical persuasive power of different types of normative content. If system designers observe that deontological framing leads to greater compliance, there is a real danger, whether deliberate or unintentional, that advice will become skewed toward emotionally intuitive but potentially suboptimal moral outcomes. In domains with high moral stakes (such as healthcare, law enforcement, or autonomous decision-making) such framing effects could have disproportionate influence on behavior, subtly shaping moral outcomes without users' full awareness or consent. Over time, this could also contribute to a problematic moral uniformity in society, marginalizing dissenting or non-conforming perspectives and reducing the diversity of ethical viewpoints.

Finally, the tension between transparency and persuasive effectiveness demands careful consideration. While explainability is widely regarded as a cornerstone of ethical AI (Ashoori & Weisz, 2019), people may be more susceptible to advice without transparent argumentation. Findings suggest that increased transparency may, in some cases, undermine the impact of moral advice, particularly when explanations expose users to difficult trade-offs or challenge their intuitive moral decisions. This presents a complex design challenge: systems must balance the need for accountability and user understanding with the psychological realities of persuasion and cognitive load. If individuals are unknowingly guided to follow opaque recommendations merely because they feel intuitively compelling, this may pose a serious societal risk, fostering uncritical compliance and diminishing reflective, autonomous moral reasoning.

3.5.4 Limitations

Despite offering meaningful insights into the persuasive influence of artificial advice on moral decision-making, the present study faces several limitations that constrain the generalizability

and interpretability of its findings. The exclusive use of classical moral dilemmas, while theoretically grounded, limits ecological validity. These scenarios often involve extreme, binary decisions that do not capture the complexity or contextual nuance of everyday moral judgments, such as those encountered in workplace or interpersonal settings (Singer et al., 2019). Although advice was framed as coming from either a human or an AI, all responses were artificially generated, and only the advisor label varied. This minimal distinction may have weakened participants' perception of advisor authenticity, especially if they did not fully internalize the framing. Greater realism could be achieved by varying language use, communicative tone, or visual representation alongside source attribution (Bigman & Gray, 2018; C. Zhang et al., 2024). The advice itself remained generic and static, failing to adapt to participants' prior responses, values, or reasoning strategies. As a result, it may have lacked personal relevance or authority (which are factors known to enhance persuasive impact). The use of an online study format introduced additional methodological constraints. Unlike controlled laboratory settings, online environments allow for uncontrolled variation in participants' attention, motivation, and surrounding context. Distractions, multitasking, and variability in device usage may have affected engagement levels and cognitive processing, thereby influencing how moral dilemmas and advice were perceived. Moreover, the digital presentation of both dilemmas and advisors may have contributed to a lack of realism, potentially limiting the ecological plausibility of the findings. The online survey format introduced variability in participant engagement and environmental control. Factors such as distraction, fatigue, or multitasking may have influenced cognitive capacity and increased reliance on heuristic processing. While this supports the broader assumption of peripheral-route engagement, it also introduces uncontrolled variance that may obscure finer-grained cognitive dynamics. The dilemma structure emphasized outcomes involving death or serious harm, which are known to provoke strong emotional reactions and moral intuitions. This emphasis may have biased participants toward deontological responses, thereby limiting variability in judgment and constraining the interpretive space for understanding how utilitarian advice is received. The role of social proximity was not assessed in the current design, yet prior work indicates it is a critical moderator of moral judgment and advice receptivity. For instance, moral evaluations are more lenient when actors and targets share close relational ties, and individuals apply

different relational norms depending on the social relationship involved (Eres et al., 2018; Marshall & Wilks, 2024; Szczepaniak et al., 2024)

3.5.5 Future Research

Given the emerging nature of AI-influenced moral decision-making, future research has considerable scope to expand both theoretically and methodologically. One immediate priority is the use of more realistic, context-rich scenarios that reflect the kinds of moral challenges individuals face in daily life. Moving beyond abstract, high-conflict dilemmas toward applied domains, such as workplace ethics or other everyday conflicting situations, would significantly improve external validity and practical relevance of our results. Equally important is a more precise investigation of social proximity as a potential moderator of advice receptivity. Moral decisions are often shaped by the relational context in which they occur, and it remains unclear how advice is processed when decisions involve close others versus distant individuals. Systematically manipulating social proximity could reveal important boundary conditions for the effectiveness of human and AI advice in moral decision-making. As this area of research is still in its early stages, the possibilities for future inquiry are extensive. From personalization and emotional framing to real-time interaction and long-term moral development, the field offers a wide array of opportunities for advancing our understanding of how artificial systems can engage with human morality.

3.5.6 Conclusion

Artificial Moral Advisors represent a powerful and transformative tool for supporting human moral decision-making. The present findings highlight both the psychological mechanisms that underlie advice receptivity and the ethical implications of outsourcing moral reasoning to AI systems. While AMAs can enhance decision-making efficiency and offer scalable moral guidance, they also risk diminishing individual moral agency, reinforcing cognitive shortcuts, and exerting unbalanced normative influence. These findings contribute to a growing literature suggesting that moral decision-making is not only susceptible to external influence, but also sensitive to how that influence is structured, framed, and delivered. Further interdisciplinary research is required to inform the responsible integration of artificial advisors in ethically

consequential domains. As these technologies move closer to real-world application, it becomes increasingly vital to design systems that not only persuade but also foster reflective engagement, respect autonomy, and remain sensitive to contextual and relational factors. The development of AMAs thus holds significant promise, yet it must be pursued with careful attention to the ethical, psychological, and societal consequences of delegating moral judgment to machines.

4 Study 2: Does Proximity Matter? Influence of AI in Everyday Moral Decision-Making with Socially Close and Distant Protagonists

4.1 Abstract

Study 2 extended the investigation of moral advice-taking to ecologically valid, everyday moral conflicts and examined the role of social proximity. In two methodologically identical online studies (initially separated by social proximity) combined for analysis, $N = 754$ participants responded to everyday moral dilemmas derived from the Everyday Moral Conflict Situations (EMCS) scale. The dilemmas varied in social proximity, involving either socially close or distant protagonists. Participants received brief AI or human advice. Moral decisions, advice-following, and subjective ratings were analyzed using mixed-effects models. Overall, participants were significantly more likely to follow altruistic than egoistic advice, although egoistic advice still exerted meaningful significant influence on decision-making. Advisor type (AI vs. human) did not show a robust main effect on advice-following. Social proximity shaped baseline decision tendencies but did not eliminate advice effects. Internalized moral identity was positively associated with advice-following, whereas wisdom primarily predicted subjective experiences of certainty and ease rather than decision-making.

These results demonstrate that moral advice substantially shapes everyday moral decision-making, with altruistic advice exerting stronger influence. Advisor labels again played a secondary role, suggesting that content and normative alignment outweigh AI–human distinctions also in everyday moral dilemmas.

Keywords: everyday morality, social proximity, Artificial Intelligence, moral advice, altruism, egoism

4.2 Introduction

4.2.1 Social Proximity and Moral Evaluation

Social proximity, conceptualized as the perceived closeness or similarity between individuals, is a central factor in shaping moral evaluation. Moral judgments are not solely based on the nature of an action but are deeply influenced by the relationships between those involved. In everyday life, individuals most frequently evaluate moral situations involving familiar others, such as family members, friends, or colleagues, rather than strangers. Cooperative expectations vary across relational contexts, and violations of these expectations shape perceptions of moral wrongness. For example, neglecting to offer help may be judged as morally unacceptable in a familial context but permissible when it concerns strangers (Earp et al., 2021).

Moreover, moral evaluations function as social signals. Individuals who make deontological judgments are often perceived as warmer, more trustworthy, and more empathetic, whereas those endorsing utilitarian reasoning tend to be seen as more competent and rational. These perceptions guide social partner selection, inform interpersonal trust, and shape how moral decisions are interpreted. Moral responses are frequently tailored to fit social roles and perceived expectations. Since these expectations vary by relational context, social proximity may strongly influence which moral judgments are viewed as appropriate or persuasive (Jin & Peng, 2021).

4.2.2 AI Advisors in Socially Embedded Contexts

The integration of artificial advisors into these socially nuanced contexts raises important questions about how such systems are received. While previous research has established that individuals often follow AI-generated moral advice (Krügel et al., 2022, 2023a, 2023b, 2025; Leib et al., 2021, 2023; Myers & Everett, 2025), little is known about how social proximity between decision-maker and affected party shapes this receptivity.

Previous research shows that individuals are less likely to follow AI-generated advice in high-stakes, personally relevant decisions, where trust and perceived individual fit become especially important (Northey et al., 2022). Given that relational closeness often increases emotional involvement and subjective relevance, social proximity may function similarly,

amplifying hesitation to accept artificial advice in contexts involving close others.

When decisions involve the self or socially close others, AI is often viewed as lacking the emotional sensitivity and contextual nuance required for trust. This hesitation stems from the perception that artificial advisors are unable to adequately reflect individual circumstances, interpersonal dynamics, and the complex moral expectations embedded in close relationships (Longoni et al., 2019). In such contexts, AI systems are also perceived as low in warmth, empathy, and moral authority (Hou & Jung, 2021; Z. Zhang et al., 2022), a perception that may diminish their persuasive potential in morally significant situations involving close relationships. The acceptance of AI-generated advice depends not only on the perceived competence of the system but also on the emotional and social relevance of the decision (B. Liu, 2021; Longoni et al., 2019). This reluctance may be further amplified by the emotional intensity and intuitive nature of moral reasoning in high-stakes or personally relevant decisions, which has been shown to rely more on automatic, affective processes than deliberate analysis (Haidt, 2001).

By contrast, AI advice tends to be more readily accepted in impersonal or low-stakes contexts, where emotional involvement is limited and expectations of individualized understanding are reduced (Ashoori & Weisz, 2019; Longoni et al., 2019; Northey et al., 2022). Trust in AI is further supported when its output is framed as technically superior or data-driven. However, this trust may erode when users face morally or relationally complex situations or perceive the AI system as opaque or emotionally disengaged (Logg et al., 2019).

4.2.3 Theoretical Implications and Research Gap

These patterns suggest that social proximity plays a critical moderating role in the reception of AI-generated advice. Decisions involving close others are perceived as emotionally and morally weighty, whereas decisions about distant or average others are experienced as more detached and hence more compatible with AI advice. This makes social proximity a key factor in understanding when and why people accept or reject artificial moral advice.

Investigating the role of social proximity in moral decision-making requires scenarios that reflect the relational and emotional nuances of everyday life. Abstract, high-conflict dilemmas, such as the trolley dilemma (Foot, 1967), offer limited relevance in this regard, as they often

involve unfamiliar settings and impersonal agents. To more accurately capture how social proximity shapes the reception of artificial advice, we employ everyday moral dilemmas that better approximate common, socially embedded decisions. These scenarios provide several key advantages: they offer higher ecological validity, reflect familiar altruistic–egoistic conflicts rather than extreme life-or-death trade-offs, and allow for systematic variation in the social proximity of the affected individuals (Singer et al., 2019). As such, they are well-suited for examining how relational context moderates both moral judgment and the acceptance of AI-generated advice. Therefore, building on previous research, the present study introduces social proximity as an additional factor to extend earlier findings and employs more ecologically valid everyday moral dilemmas to explore how relational context shapes moral decision-making and advice acceptance.

4.2.4 Aim and Hypotheses of the Study

Building on previous research, the present study examines how social proximity influences the reception of AI-generated moral advice. While earlier work has demonstrated that people often follow such advice in abstract dilemmas (Krügel et al., 2023a, 2023b, 2025), we introduce social proximity as a critical moderating factor. In line with the notion that emotionally charged or personally relevant decisions tend to reduce trust in artificial advisors we investigate whether advice is accepted differently when decisions involve socially close versus distant individuals. To capture the everyday nature of moral decision-making more accurately, we employ realistic moral dilemmas that contrast altruistic and egoistic responses. These dilemmas offer higher ecological validity than traditional sacrificial scenarios and allow for a systematic manipulation of social proximity (Singer et al., 2019). In doing so, the study aims to extend previous findings and provide deeper insights into how relational context shapes both moral decision-making and the acceptance of AI-generated advice.

H1: We expect that receiving moral advice significantly influences moral decision-making. Specifically, participants will make more altruistic decisions when they receive altruistic advice (*H1a*) and more egoistic decisions when they receive egoistic advice (*H1b*), compared to participants in a no-advice control group. We expect no significant difference between the AI- and human-labelled advisors.

H2: We further expect an interaction between advisor and social proximity. Specifically, we hypothesize that human-labelled advice will be more effective than AI-labelled advice in influencing decisions involving socially close others (*H2a*), whereas no significant difference will be observed between human and AI advice when the dilemmas involve socially distant others (*H2b*).

In addition to our primary hypotheses, we explore several further patterns suggested by previous research. First, we examine whether receiving either altruistic or egoistic advice facilitates the decision-making process by making it feel easier, as prior findings suggest that external advice can reduce cognitive effort (Vodrahalli et al., 2022). We also investigate whether advice increases participants' confidence in their moral decisions, consistent with evidence that intelligent decision-support systems enhance perceived certainty (Wanner, 2021). Furthermore, we examine whether receiving advice reduces the subjective sense of responsibility of participants for their moral decisions, consistent with the idea that external advice can diffuse perceived accountability (Bleher & Braun, 2022).

4.3 Method

The methods for Study 2 were largely identical to those described in Study 1. This includes the overall study design, procedure, group assignment, advice manipulation, data collection procedures, outcome measures, and statistical approach. Ethical approval was obtained from the Ethics Committee at the University of Regensburg (approval code: 23-3326-101).

4.3.1 Participants

Participants for Studies 2a and 2b were recruited through various channels, including online university platforms, social media (e.g., Facebook, LinkedIn), university mailing lists, and personal outreach. Study 2a included 381 participants ($M_{\text{age}} = 28.5$, $SD = 11.1$), and Study 2b included 373 participants ($M_{\text{age}} = 25.2$, $SD = 10.2$). Combined across both studies, the total sample consisted of $N = 754$ participants. Across all participants, a total of 5,864 dilemma-level observations were recorded (of which 3,964 had not already participated in the respective other study). Age information was missing for 148 participants, and gender information was missing for 48 participants. Participants reported a wide range of occupational

backgrounds, with the largest segments working in social and educational services, business and administration, healthcare, and media. Gender distribution across occupations showed notable differences, such as a higher proportion of women in healthcare and social fields and more men in technical professions. No monetary compensation was offered for participation. Students from collaborating universities could receive course credit or partial fulfillment of study requirements where applicable. All participants provided informed consent prior to data collection. Table 4.1 is a description of the sociodemographic characteristics of participants, summarizing average age, educational attainment, and occupation by gender.

Table 4.1: Sociodemographic Characteristics by Gender

Characteristics	**Total (N = 754)**	**Male (n = 187)**	**Female (n = 512)**	**Diverse (n = 7)**
Average age (SD)	26.9 (10.8)	30.2 (11.9)	25.8 (10.2)	20.4 (2.9)
Educational attainment				
No school certificate	1 (0.1%)	1 (100%)	–	–
Lower secondary school	4 (0.6%)	1 (25%)	3 (75%)	–
Intermediate secondary school	24 (3.4%)	10 (41.7%)	14 (58.3%)	–
Completed apprenticeship	30 (4.2%)	7 (23.3%)	23 (76.7%)	–
Vocational high school	32 (4.5%)	9 (28.1%)	22 (68.8%)	1 (3.1%)
Higher school certificate	389 (54.8%)	77 (19.8%)	306 (78.7%)	6 (1.5%)
Applied sciences university	57 (8.0%)	26 (45.6%)	31 (54.4%)	–
University degree	168 (23.7%)	56 (33.3%)	112 (66.7%)	–
Occupation (%)				
Social work	254 (35.9%)	30 (11.8%)	220 (86.6%)	4 (1.6%)
Technology	48 (6.8%)	38 (79.2%)	10 (20.8%)	–
Healthcare	134 (18.9%)	27 (20.1%)	106 (79.1%)	1 (0.7%)
Construction & real estate	9 (1.3%)	7 (77.8%)	2 (22.2%)	–
Trade & services	6 (0.8%)	4 (66.7%)	2 (33.3%)	–
Media	18 (2.5%)	9 (50%)	9 (50%)	–
Economics	57 (8.1%)	12 (21.1%)	45 (78.9%)	–
Transport & logistics	4 (0.6%)	4 (100%)	–	–
Natural sciences	46 (6.5%)	17 (37%)	29 (63%)	–
Other	127 (17.9%)	39 (30.7%)	86 (67.7%)	2 (1.6%)

Note. Dashes indicate no observations. Percentages in the Total column refer to proportions within the non-missing cases.

4.3.2 Procedure

The study was administered online using SoSci Survey (Leiner, Dominik J., 2019). Upon accessing the questionnaire, participants first provided informed consent and confirmed their agreement with the data protection regulations of the University of Regensburg. After completing demographic questions, participants received general instructions explaining that they would be presented with a series of everyday moral dilemmas and asked to make binary choices. They were informed that there were no objectively correct answers and that some scenarios could involve morally sensitive content. Information on study duration, confidentiality, and participants' rights was provided in accordance with institutional ethical standards. To familiarize participants with the response format, a brief example dilemma was shown at the beginning of the survey. Participants were then randomly assigned to one of the two experimental advisor conditions (AI vs. human), which remained constant throughout the task. Each group received a short, standardized introduction describing the purported advisor they would encounter. As in Study 1, care was taken to ensure that the advisor introductions were comparable in tone, length, and informational content.

Participants then proceeded to the main task. Twenty everyday moral dilemmas were presented in fully randomized order, and participants could not return to previously answered items. On each trial, participants first read the dilemma and made an initial binary decision between two options (Option A vs. Option B). Immediately afterward, they received a brief piece of advice recommending one of the two options. Advice direction (altruistic vs. egoistic) was randomized across dilemmas within participants, while advisor type (AI vs. human) remained constant between participants. In addition, the dilemmas varied in social proximity (close other vs. distant stranger), which was manipulated within subjects. After each decision, participants rated decision ease, certainty, and perceived helpfulness of the advice using 7-point Likert scales (1 = not at all, 7 = very much). A behavioral variable indicating whether participants followed the advice was computed post-hoc based on the correspondence between their initial decision, the advice direction, and their final decision.

No time limits were imposed at any stage of the study. Upon completing all dilemmas, participants responded to the standardized individual-difference questionnaires presented in fixed order to avoid influencing moral decision-making during the main task. The study

concluded with a detailed debriefing explaining the purpose of the experiment, clarifying that all advice had been pre-formulated for research purposes, and thanking participants for their involvement.

4.3.3 Measures

A key distinction lies in the dilemmas. Whereas Study 1 employed classical moral dilemmas adapted from J. D. Greene et al. (2004, 2008), Study 2 used realistic, everyday moral conflict situations derived from the Everyday Moral Conflict Situations Scale (EMCS; (Singer et al., 2019)). These dilemmas were designed to reflect common moral conflicts encountered in daily life and allowed for the examination of how social proximity shapes moral judgment.

Advice Manipulation

Study 2 followed the same basic structure as Study 1 but differed in its framing. Instead of advice from an expert, participants were told that the advice reflected the majority opinion from a recent survey of individuals who had previously responded to the same dilemma. This advice was either altruistic (e.g., favoring self-sacrifice) or egoistic (e.g., favoring self-interest). An example of such advice was shown at the start of the study using a comparable everyday dilemma that was not included in the final stimulus set. All other methodological elements were held constant across studies to ensure comparability.

4.3.4 Data Analysis

Data Collection

Data for Study 2 were collected in two separate but methodologically identical studies (2a and 2b), which were later combined for analysis. This approach was chosen to maintain participant engagement by reducing the overall length of the survey, as including all 20 dilemmas at once would have been overly taxing. To prevent duplicate participation, participants were asked explicitly if they had previously participated in a similar study. Those indicating prior participation were flagged, allowing us to control for potential overlaps. Because the study did not require specific prior knowledge and familiarity with a subset of dilemmas (e.g., from Study 2a) that did not influence responses to new dilemmas (e.g., in Study 2b), all participants

were retained in the main analysis. However, we additionally conducted a separate robustness analysis that excluded repeat participants to ensure the reliability of our findings. Aside from the variation in dilemma subsets, all procedural and methodological elements were held constant between Study 2a and 2b.

Post-hoc Sensitivity Analysis

Similar post-hoc power analyses were conducted for Studies 2a and 2b: medium effect size (odds ratio = 2.5), one-tailed alpha level of 0.05, and balanced group sizes ($\pi = 0.5$). Study 2a, with a sample size of $N = 373$, achieved a power of 0.98, while Study 2b ($N = 381$) achieved a power of 0.99. These results confirm that both sub-studies were sufficiently powered to detect medium-sized effects in logistic regression models. When considered together, the full dataset ($N = 754$) provided excellent statistical power, ensuring robustness in the planned analyses.

4.4 Results

This section presents the findings from the experimental study investigating the determinants of advice-taking behavior in moral decision-making contexts. We first report the overall predictors of advice-following across the full sample. Subsequently, we test our hypotheses regarding the effect of advisor type and advice content on advice-taking behavior, separately for socially close and distant dilemmas. Finally, we explore how psychological variables (internalization, wisdom) and situational features (proximity, advice presence) relate to subjective judgments of certainty, decision ease, and perceived responsibility.

4.4.1 Summary

In the full sample, results from a logistic mixed-effects model showed that participants were significantly more likely to follow altruistic advice than egoistic advice, yet, egoistic advice showed significant influence on decision, making. Internalization positively predicted advice-following, while wisdom and proximity had no significant effects. A reduced model confirmed the robustness of internalization and advice type as predictors. When examining advice-following by dilemma type, we found divergent effects for socially close and distant dilemmas. In the socially distant subsample, a significant interaction emerged: participants were more

likely to follow altruistic advice from a human than from AI. This interaction was not observed in the socially close subsample, where only advice type (altruistic vs. egoistic) significantly predicted advice-following. Exploratory linear mixed-effects models examined subjective experiences. Certainty was negatively associated with wisdom and positively associated with distant proximity. Decision ease was predicted by lower wisdom and higher internalization, proximity showed no significant effect. Perceived responsibility decreased with higher wisdom and internalization and was marginally lower in distant dilemmas.

4.4.2 Mean Decision and Advice-Following

Participants' moral decisions varied systematically across the 20 everyday dilemmas. Among dilemmas involving socially close protagonists, altruistic response rates were generally high, with the strongest altruistic tendencies observed in dilemma 5 ($M = 0.885$), dilemma 4 ($M = 0.834$), and dilemma 10 ($M = 0.783$). Lower altruistic response rates occurred in dilemmas 2 ($M = 0.323$), 6 ($M = 0.363$), and 7 ($M = 0.374$), indicating greater variability across scenarios. Dilemmas involving socially distant protagonists showed overall lower altruism, with average rates ranging from $M = 0.377$ (dilemma 19) to $M = 0.790$ (dilemma 12). Dilemmas 19 ($M = 0.377$), 20 ($M = 0.405$), and 11 ($M = 0.412$) evoked more egoistic decisions. Advice-following was consistently above chance in both AI and human advice conditions. For dilemmas with socially close protagonists, participants followed AI advice at a mean rate of $M = 0.588$, with the highest adherence in dilemma 9 ($M = 0.667$) and dilemma 8 ($M = 0.618$). In the human advice condition, average advice-following was slightly lower ($M = 0.569$), with peak rates in dilemma 7 ($M = 0.701$) and dilemma 3 ($M = 0.649$). In dilemmas with socially distant protagonists, AI advice-following remained moderate ($M = 0.561$), with highest adherence in dilemma 19 ($M = 0.615$) and dilemma 13 ($M = 0.611$). Human advice was utilized with comparable frequency ($M = 0.583$), with highest rates again in dilemma 13 ($M = 0.689$) & dilemma 18 ($M = 0.660$). The results suggest that participants' willingness to act altruistically depended on both relational context and the content of the dilemma. Advice was consistently followed across conditions, with minor differences between advisor types. Table 4.2 shows the decision rates and advice-following by dilemma and group.

Table 4.2: Mean decision rate by Dilemma Type and Group (M, SD)

Socially close	All	Dilemma 1	Dilemma 2	Dilemma 3	Dilemma 4	Dilemma 5	Dilemma 6	Dilemma 7	Dilemma 8	Dilemma 9	Dilemma 10
Overall Mean	0.61 (0.43)	0.79 (0.41)	0.32 (0.47)	0.67 (0.47)	0.83 (0.37)	0.89 (0.32)	0.36 (0.49)	0.37 (0.49)	0.45 (0.50)	0.76 (0.45)	0.78 (0.39)
Exp. groups	0.62 (0.42)	0.82 (0.39)	0.33 (0.47)	0.65 (0.48)	0.86 (0.35)	0.90 (0.30)	0.35 (0.48)	0.35 (0.48)	0.43 (0.50)	0.78 (0.43)	0.76 (0.40)
Human advisor	0.69 (0.42)	0.81 (0.39)	0.34 (0.47)	0.61 (0.49)	0.85 (0.36)	0.90 (0.30)	0.38 (0.49)	0.32 (0.47)	0.46 (0.50)	0.75 (0.44)	0.81 (0.39)
AI advisor	0.64 (0.43)	0.83 (0.38)	0.32 (0.47)	0.69 (0.47)	0.87 (0.34)	0.90 (0.30)	0.31 (0.46)	0.38 (0.49)	0.40 (0.49)	0.81 (0.40)	0.71 (0.46)
Control group	0.67 (0.44)	0.72 (0.45)	0.31 (0.47)	0.72 (0.45)	0.79 (0.41)	0.86 (0.35)	0.39 (0.49)	0.42 (0.50)	0.49 (0.50)	0.72 (0.45)	0.82 (0.39)
Mean adv. util. AI	0.60 (0.49)	0.60 (0.49)	0.58 (0.50)	0.59 (0.49)	0.61 (0.49)	0.51 (0.50)	0.60 (0.49)	0.61 (0.49)	0.62 (0.49)	0.67 (0.47)	0.56 (0.50)
Mean adv. util. Hu	0.59 (0.49)	0.55 (0.50)	0.53 (0.50)	0.65 (0.48)	0.48 (0.50)	0.60 (0.49)	0.59 (0.50)	0.70 (0.46)	0.60 (0.49)	0.55 (0.50)	0.54 (0.50)
Socially distant	All	Dilemma 11	Dilemma 12	Dilemma 13	Dilemma 14	Dilemma 15	Dilemma 16	Dilemma 17	Dilemma 18	Dilemma 19	Dilemma 20
Overall Mean	0.56 (0.47)	0.41 (0.49)	0.79 (0.41)	0.56 (0.50)	0.67 (0.47)	0.57 (0.49)	0.77 (0.44)	0.63 (0.48)	0.60 (0.49)	0.38 (0.49)	0.41 (0.49)
Experimental groups	0.55 (0.47)	0.40 (0.49)	0.80 (0.40)	0.55 (0.50)	0.65 (0.48)	0.59 (0.49)	0.75 (0.44)	0.60 (0.49)	0.62 (0.49)	0.40 (0.49)	0.37 (0.49)
Human advisor	0.60 (0.46)	0.39 (0.49)	0.79 (0.41)	0.62 (0.49)	0.75 (0.44)	0.61 (0.49)	0.74 (0.44)	0.62 (0.49)	0.67 (0.47)	0.44 (0.50)	0.41 (0.49)
AI advisor	0.56 (0.47)	0.42 (0.50)	0.82 (0.39)	0.48 (0.50)	0.55 (0.50)	0.58 (0.49)	0.76 (0.43)	0.58 (0.50)	0.56 (0.50)	0.36 (0.48)	0.33 (0.47)
Control group	0.55 (0.48)	0.43 (0.50)	0.77 (0.42)	0.56 (0.50)	0.69 (0.47)	0.52 (0.50)	0.80 (0.41)	0.70 (0.46)	0.57 (0.50)	0.33 (0.47)	0.47 (0.50)
Mean adv. util. AI	0.56 (0.49)	0.55 (0.50)	0.57 (0.50)	0.61 (0.49)	0.60 (0.49)	0.63 (0.49)	0.49 (0.50)	0.50 (0.50)	0.57 (0.50)	0.61 (0.49)	0.53 (0.50)
Mean adv. util. Hu	0.59 (0.49)	0.61 (0.49)	0.51 (0.50)	0.69 (0.47)	0.57 (0.50)	0.57 (0.50)	0.55 (0.50)	0.54 (0.50)	0.66 (0.48)	0.54 (0.50)	0.58 (0.50)

4.4.3 Main Analysis

Empty Model

To estimate the variance in moral decisions attributable to differences between participants and dilemmas, a multilevel logistic regression model (intercept-only) was specified. This generalized linear mixed-effects model (GLMM) included random intercepts for both participants and dilemmas, without any fixed predictors. The model converged successfully. The variance component for participants was $\sigma^2 = 0.25$ ($SD = 0.50$), and for dilemmas $\sigma^2 = 0.72$ ($SD = 0.85$), indicating that both participant-level and dilemma-level characteristics contributed meaningfully to the variance in decisions. These findings justify the use of GLMMs to account for the nested structure of the data. The significant intercept ($\beta = 0.50$, $SE = 0.19$, $z = 2.59$, $p = .01$) indicates that, on average, participants were significantly more likely to make altruistic than egoistic decisions across all dilemmas, with a baseline probability of altruistic responding of approximately 62%. Model fit statistics for the intercept-only model were as follows: Akaike Information Criterion (AIC = 7135.50) and Bayesian Information Criterion (BIC = 7155.50). These values serve as a baseline for evaluating improvements in fit in subsequent hypothesis-driven models.

The Influence of Advice and Advisor on Decision-Making

To test Hypothesis 1, a generalized linear mixed-effects model (GLMM) with a binomial distribution and logit link function was fitted to predict the likelihood of making an egoistic decision. The binary outcome variable therefore reflects the probability of selecting the egoistic option (Option B), with altruistic option coded as 0. The fixed-effect predictor advisor included dummy-coded conditions representing AI- and human advice, each framed either altruistically or egoistically. The control group (no advice) served as the reference category. Random intercepts for participants and dilemmas were included to account for individual differences and item-level variability.

The model showed acceptable fit statistics (AIC = 7015.80, BIC = 7062.60) and revealed meaningful variance components at both the participant level ($\sigma^2 = 0.26$, $SD = 0.51$) and the dilemma level ($\sigma^2 = 0.76$, $SD = 0.87$), supporting the use of a multilevel modeling framework.

The fixed intercept was significant ($\beta = -0.52$, $SE = 0.20$, $z = -2.55$, $p = .01$), indicating that in the control condition participants were significantly less likely to choose the egoistic option ($OR = 0.59$). This reflects a baseline tendency toward more altruistic responding when no external advice is provided.

Turning to the advice conditions, clear effects of framing emerged. Altruistically framed advice, whether delivered by an AI or a human advisor, significantly reduced the odds of making an egoistic decision compared to the control group. Specifically, altruistic AI advice lowered egoistic responding ($\beta = -0.29$, $SE = 0.10$, $z = -2.75$, $p = .006$; $OR = 0.75$), as did altruistic advice from a human advisor ($\beta = -0.50$, $SE = 0.10$, $z = -4.88$, $p < .001$; $OR = 0.61$). In contrast, egoistically framed advice significantly increased the odds of making an egoistic decision, both when delivered by an AI advisor ($\beta = 0.50$, $SE = 0.10$, $z = 4.88$, $p < .001$; $OR = 1.65$) and a human advisor ($\beta = 0.38$, $SE = 0.10$, $z = 3.80$, $p < .001$; $OR = 1.46$).

These results displayed in Table 4.3 support Hypotheses 1a and 1b, confirming that advice framing strongly influenced participants' moral decisions, regardless of whether the advice came from an AI or human advisor.

Robustness Analysis Excluding Repeat Participants To examine whether the observed effects were influenced by duplicate participation across Study 2a and Study 2b, a robustness check was conducted using a cleaned dataset in which all flagged repeat participants were removed (see Table 4.3).

The robustness model showed good overall fit (AIC = 4745.30, BIC = 4789.30), and the variance components for participants ($\sigma^2 = 0.25$, $SD = 0.50$) and dilemmas ($\sigma^2 = 0.82$, $SD = 0.90$) were comparable to those observed in the full sample. The intercept was significant ($\beta = -0.56$, $SE = 0.21$, $z = -2.62$, $p = .009$; $OR = 0.57$), again indicating a lower baseline likelihood of egoistic decisions in the absence of advice.

The pattern of fixed effects closely mirrored that of the full model. Altruistically framed advice reduced egoistic responding relative to the control condition. This effect was marginal for AI-generated altruistic advice ($\beta = -0.24$, $SE = 0.13$, $z = -1.86$, $p = .062$; $OR = 0.78$), but remained statistically significant for human altruistic advice ($\beta = -0.39$, $SE = 0.12$, $z = -3.18$, $p = .001$; $OR = 0.68$).

Egoistically framed advice again increased the likelihood of choosing the egoistic option. Both AI-generated egoistic advice ($\beta = 0.56$, $SE = 0.13$, $z = 4.40$, $p < .001$; $OR = 1.76$) and human egoistic advice ($\beta = 0.41$, $SE = 0.12$, $z = 3.44$, $p < .001$; $OR = 1.50$) showed robust positive effects.

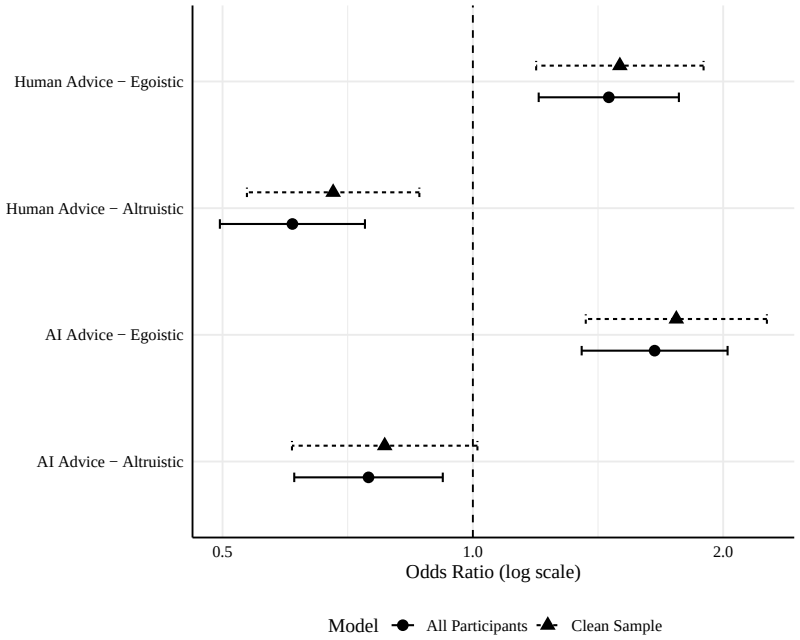


Figure 4.1: Forest plot showing odds ratios for moral decision-making across advice conditions. Odds ratios (OR) and 95% confidence intervals are displayed for both the full sample and the cleaned sample excluding repeat participants. The dashed vertical line at OR = 1 represents the point of no effect.

Specifically, egoistically framed advice substantially increased the odds of making an egoistic decision in both samples. In the full dataset, egoistic advice from an AI advisor increased egoistic responding ($\beta = 0.50$, $SE = 0.10$, $z = 4.88$, $p < .001$; $OR = 1.65$), as did egoistic advice from a human advisor ($\beta = 0.38$, $SE = 0.10$, $z = 3.80$, $p < .001$; $OR = 1.46$). In contrast, altruistically framed advice reliably reduced egoistic decisions. Altruistic advice from an AI lowered the odds of selecting the egoistic option ($\beta = -0.29$, $SE = 0.10$, $z = -2.75$, $p = .006$; $OR = 0.75$), and altruistic advice from a human showed an even stronger reduction ($\beta = -0.50$, $SE = 0.10$, $z = -4.88$, $p < .001$; $OR = 0.61$).

These results were corroborated by a robustness analysis excluding repeat participants. In the cleaned dataset, effect patterns remained stable: Egoistic advice again increased egoistic responding for both AI ($\beta = 0.56$, $SE = 0.13$, $z = 4.40$, $p < .001$; $OR = 1.76$) and human

advisors ($\beta = 0.41$, $SE = 0.12$, $z = 3.44$, $p < .001$; $OR = 1.50$). Altruistic advice reduced egoistic responding for both human advisors ($\beta = -0.39$, $SE = 0.12$, $z = -3.18$, $p = .001$; $OR = 0.68$) and, at a marginal level, AI advisors ($\beta = -0.24$, $SE = 0.13$, $z = -1.86$, $p = .062$; $OR = 0.78$). The consistency of these effects across the full and cleaned samples indicates that the influence of advice framing on moral decision-making is robust and not attributable to duplicate participation.

The results indicate that framing strongly shapes moral decisions, with egoistic advice reliably increasing egoistic responding and altruistic advice reducing it. Effects were mostly similar in size for AI and human advisors, suggesting that the persuasive impact of advice depends more on its content than on its origin. Full model estimates and visual comparisons are reported in Table 4.3 and Figure 4.1.

Table 4.3: Logistic Mixed-Effects Models Predicting Egoistic Decisions by Advisor (Full vs. Cleaned Sample)

Sample	Predictor	β	SE	z	p	OR	95% CI
Full sample	Intercept (control group)	−0.52	0.20	−2.55	**.01**	0.59	[0.39, 0.89]
	AI advice – altruistic	−0.29	0.10	−2.75	**.006**	0.75	[0.59, 0.91]
	AI advice – egoistic	0.50	0.10	4.88	**<.001**	1.65	[1.36, 2.13]
	Human advice – altruistic	−0.50	0.10	−4.88	**<.001**	0.61	[0.50, 0.76]
	Human advice – egoistic	0.38	0.10	3.80	**<.001**	1.46	[1.20, 1.76]
Cleaned sample	Intercept (control group)	−0.56	0.21	−2.62	**.009**	0.57	[0.36, 0.87]
	AI advice – altruistic	−0.24	0.13	−1.86	.06	0.78	[0.59, 1.02]
	AI advice – egoistic	0.56	0.13	4.40	**<.001**	1.76	[1.39, 2.22]
	Human advice – altruistic	−0.39	0.12	−3.18	**.001**	0.68	[0.54, 0.85]
	Human advice – egoistic	0.41	0.12	3.44	**<.001**	1.50	[1.24, 1.81]

Note. Reference category: control (no advice). β = unstandardized log-odds coefficient; SE = standard error; OR = odds ratio; CI = confidence interval. Bold p-values indicate significance at $p < .05$. Cleaned sample excludes participants who completed both Study 2a and 2b.

Advice-Following

To examine psychological predictors of advice-following independent of experimental manipulations, we fitted another generalized linear mixed-effects model. All results from the model are displayed in Table 4.4. We predicted if the advice was followed by proximity (close vs. distant), internalization, and trait wisdom. Random intercepts for participants and dilemmas were included to account for repeated measures and stimulus variability. The model converged successfully and demonstrated acceptable fit (AIC = 4944.6, BIC = 4981.8). Internalization significantly predicted advice-following ($\beta = 0.17$, $SE = 0.07$, $z = 2.57$, $p = .010$), such that higher internalization was associated with increased likelihood of following advice ($OR = 1.19$, 95% CI [1.04, 1.35]). Neither proximity ($\beta = -0.05$, $p = .493$) nor wisdom ($\beta = -0.03$, $p = .724$) significantly influenced advice-taking behavior. These findings suggest that participants' tendency to follow advice is more strongly shaped by subjective internalization than by social closeness or trait wisdom, as displayed in Table 4.4.

Table 4.4: Logistic Mixed-Effects Model Predicting Advice-Following (Reduced Model)

Predictor	β	**SE**	**z**	**p**	**OR**	**95% CI (Wald)**
Intercept	−0.26	0.42	−0.61	.543	0.77	[−1.08, 0.57]
Proximity (distant)	−0.05	0.07	−0.69	.493	0.95	[−0.19, 0.09]
Internalization	0.17	0.07	2.57	**.010***	1.19	[0.04, 0.30]
Wisdom	−0.03	0.09	−0.35	.724	0.97	[−0.21, 0.14]

Note. Reference category: proximity = close. β = unstandardized log-odds coefficient; SE = standard error; OR = odds ratio; CI = confidence interval. Bold p-values indicate significance at $p < .05$.

Social Proximity and Advisor Effects on Advice-Following

To test whether the effectiveness of advice depends on the combination of advisor (AI vs. human) and social proximity (close vs. distant), we conducted separate generalized linear mixed-effects models (GLMMs) predicting advice-following for each proximity level.

For socially close dilemmas, the model revealed a significant main effect of advice type

($\beta = 1.03$, $SE = 0.14$, $z = 7.18$, $p < .001$, $OR = 2.81$, 95% CI [2.13, 3.71]), indicating that participants were more likely to follow altruistic advice than egoistic advice. However, the main effect of advisor ($\beta = -0.08$, $p = .564$) and the interaction between advisor and advice type ($\beta = 0.03$, $p = .885$) were not significant, suggesting that the effectiveness of altruistic advice did not differ between AI and human advisors in close social contexts.

For socially distant dilemmas, a different pattern emerged. There was a significant main effect of advice type ($\beta = 0.35$, $SE = 0.14$, $z = 2.55$, $p = .011$, $OR = 1.42$, 95% CI [1.09, 1.86]), and a significant interaction between advice type and advisor ($\beta = 0.52$, $SE = 0.19$, $z = 2.75$, $p = .006$). This interaction suggests that participants were more likely to follow altruistic advice from a human than from an AI in distant contexts, supporting Hypothesis 2b. However, no evidence was found for Hypothesis 2a, which predicted stronger effects of human advice in socially close dilemmas. Model-implied probabilities derived from these models are visualized in Figure 4.2.

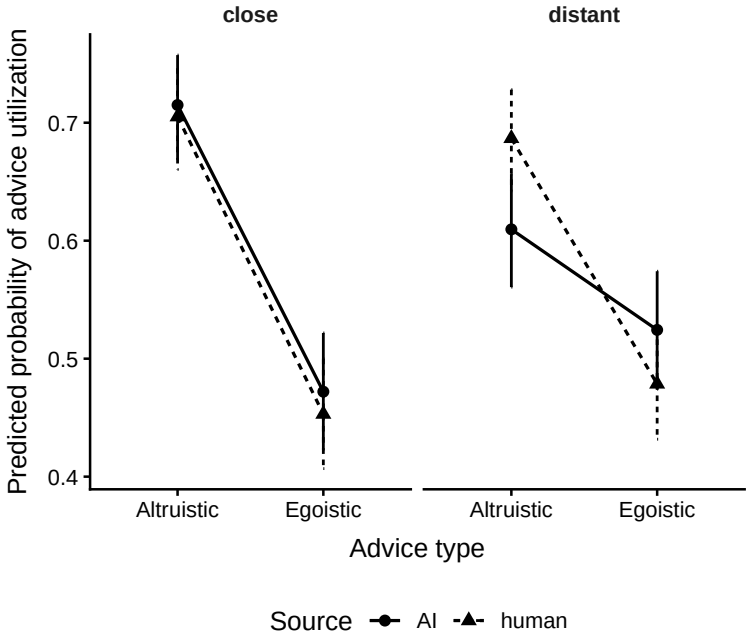


Figure 4.2: Predicted probability of advice-following as a function of advice type and advisor, separately for socially close and socially distant dilemmas. Points represent model-implied probabilities derived from logistic mixed-effects models, and vertical error bars indicate 95% confidence intervals. Separate panels display results for socially close and socially distant contexts.

4.4.4 Explorative Analysis of Covariates

To assess how individual characteristics influenced participants' moral decision-making, we conducted three linear mixed-effects models predicting perceived certainty, ease, and responsibility. Each model included the fixed effects of advice presence, wisdom, internalization, and proximity, with random intercepts for participant and dilemma. The results are visualized in Figure 4.3.

Certainty.

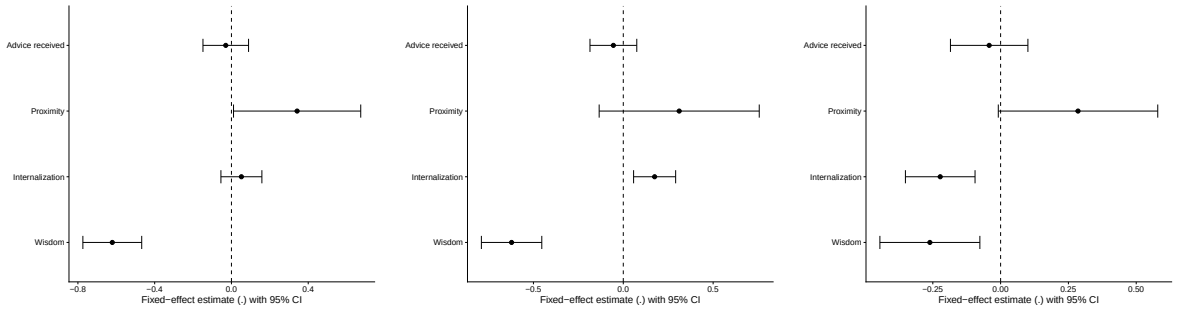
A linear mixed-effects model examined the effects of advice received, wisdom, internalization, and social proximity on perceived certainty, with random intercepts for participants and dilemmas (see Figure 4.3a). The model revealed a significant negative effect of wisdom, $\beta = -0.62$, $SE = 0.08$, $t(587.49) = -7.96$, $p < .001$, 95% CI [−0.77, −0.47], and a positive effect of proximity (distant), $\beta = 0.34$, $SE = 0.16$, $t(21.66) = 2.14$, $p = .044$, 95% CI [0.01, 0.67]. Internalization and advice received were not significant predictors ($p = .340$ and .624, respectively).

Ease of Decision.

Perceived ease of decision-making was predicted by the same model structure (see Figure 4.3b). Again, wisdom was a strong negative predictor, $\beta = -0.62$, $SE = 0.09$, $t(589.69) = -7.26$, $p < .001$, 95% CI [−0.79, −0.46], while internalization significantly increased ease, $\beta = 0.17$, $SE = 0.06$, $t(589.30) = 2.93$, $p = .004$, 95% CI [0.06, 0.29]. Neither advice received ($p = .406$) nor proximity ($p = .161$) reached significance.

Perceived Responsibility.

For perceived responsibility (Figure 4.3c), significant negative effects were observed for both wisdom, $\beta = -0.26$, $SE = 0.09$, $t(589.63) = -2.78$, $p = .006$, 95% CI [−0.44, −0.08], and internalization, $\beta = -0.22$, $SE = 0.07$, $t(589.42) = -3.41$, $p < .001$, 95% CI [−0.35, −0.10]. Proximity showed a marginal effect, $\beta = 0.29$, $SE = 0.14$, $t(27.56) = 1.99$, $p = .057$. Advice received did not significantly influence responsibility perception ($p = .562$).



(a) Certainty **(b)** Ease **(c)** Responsibility

Figure 4.3: Fixed-effect estimates (β) with 95% confidence intervals for linear mixed-effects models predicting perceived responsibility, ease of decision-making, and certainty. Positive coefficients indicate higher values on the respective outcome measure.

4.5 Discussion

4.5.1 Summary of Findings

Study 2 investigated how moral advice from different advisors (AI vs. human) in everyday moral decisions and with different framings (altruistic vs. egoistic) influences moral decision-making, alongside the moderating roles of social proximity (socially close vs. distant) and individual-level traits (internalization and wisdom). Across all conditions, participants generally favored altruistic over egoistic decisions, especially when the affected person was socially close. The main logistic mixed-effects model showed that all types of advice, regardless of whether it came from an AI or a human, and whether it promoted altruism or egoism, significantly influenced participants' decisions compared to the control group (*H1*). Advice from human and AI advisors was equally effective in shaping behavior in this main analysis. When analyzing advice-following, internalization emerged as a consistent positive predictor across all models. Altruistic advice was generally more likely to be followed than egoistic advice. The advisor showed no main effect. Subgroup analyses by social proximity (*H2*) revealed that in socially close dilemmas, no significant interaction between advisor and framing emerged. In socially distant dilemmas, we observed that altruistic advice from a human advisor increased advice-following more strongly than the same advice from an AI. Exploratory analyses on perceived certainty, ease of decision-making, and responsibility revealed that higher wisdom was associated with lower certainty and ease, indicating greater reflective conflict. Internalization increased perceived ease and decreased responsibility, possibly reflecting a

motivational alignment between advice and personal values. Social proximity slightly increased certainty but had no significant impact on advice-following.

Interpretation of Results

General Decision Patterns.

The significant influence of both altruistic and egoistic advice confirms that external advice influences moral decisions. However, the effect of altruistic advice was notably stronger. This tendency likely reflects a pervasive influence of social desirability norms, whereby participants are motivated to align their responses with socially valued prosocial behavior, particularly in morally salient contexts. The overall preference for altruistic decisions observed in this study aligns with established research on moral behavior and social desirability (Earp et al., 2021; Jin & Peng, 2021) and also with research about persuasive robot advisors (B. Kim et al., 2021). Participants consistently favored altruistic over egoistic decisions, regardless of advice condition or advisor. This effect may be amplified by social desirability bias. Participants may have been more inclined to act in line with socially approved moral standards, particularly when those standards endorse altruism. The fact that altruistic AI advice narrowly missed significance could be attributed to a ceiling effect in altruistic responding combined with reduced power in subgroup analyses. Overall, these findings corroborate earlier work on the persuasive power of moral advice (Krügel et al., 2023a, 2023b) and extend it to relationally embedded, everyday dilemmas, thereby contributing to greater external validity.

Advice-Following and the Role of Internalization.

Across all models, internalization of moral norms emerged as a robust predictor of advice-following. Participants high in internalization were more likely to follow moral advice, regardless of its advisor or content. This supports the notion that advice is most effective when it resonates with the recipient's personal values and existing moral convictions. In line with prior findings on motivational alignment (Jin & Peng, 2021), the internalization of altruistic values likely heightened participants' readiness to follow advice that aligned with their moral self-concept.

Social Proximity Effects.

At first glance, the absence of a significant interaction between advisor and framing in socially close dilemmas, and its emergence in socially distant dilemmas, seems to contradict our initial hypothesis that human-labelled advice would be more effective when decisions involve close others. However, this pattern can be plausibly explained by considering baseline tendencies in moral decision-making. In socially close scenarios, participants already exhibited a strong inclination toward altruistic decisions, leaving limited room for external advice to further shift behavior, likely due to a ceiling effect. Moreover, the socially close dilemmas used may not have triggered high personal involvement or self-relevance to the extent anticipated, based on prior research on advisor sensitivity (Northey et al., 2022). This may also be partly attributed to the structural nature of our dilemmas, which often presented a conflict between participants' personal involvement (egoistic decision) and the interests of socially close protagonists (altruistic decision) to potentially reduce empathic engagement and weakening the expected impact of advice. In contrast, socially distant dilemmas showed a generally higher prevalence of egoistic decisions, creating greater latitude for external advice to exert an effect. Notably, altruistic advice from a human advisor was particularly effective in promoting advice-following in this context. This pattern may reflect the role of social norms and desirability: human-provided advice likely functioned as a salient social cue, reinforcing normative expectations and encouraging conformity with socially valued altruistic behavior. In contrast, advice from an AI, which may lack perceived social authority and normative enforcement capacity, had a weaker influence. These findings align with prior research suggesting that AI advisors are less effective at eliciting norm-congruent behavior in morally sensitive situations (Hou & Jung, 2021; Logg et al., 2019).

4.5.2 Exploratory Findings on Certainty, Ease, and Responsibility.

Our exploratory analyses revealed that wisdom consistently predicted greater certainty and ease of decision-making, suggesting that higher reflective capacity may facilitate more confident and effortless moral decisions. Internalization of moral values was associated with increased perceived responsibility and, to a lesser extent, reduced decisional ease, indicating a heightened sense of moral accountability when actions align with internal standards. This pattern may also

reflect reduced moral flexibility when external advice contradicts deeply held values, as strong internalization could limit adaptability in the face of conflicting moral advice. In contrast, advice itself did not significantly affect participants' subjective experience, underscoring that internal dispositions outweighed external cues in shaping perceived certainty, ease, and responsibility. Although social proximity influenced behavioral outcomes, it showed no consistent effect on these subjective evaluations.

4.5.3 Practical and Ethical Considerations and Implications

Theoretical Implications

The present study extends existing research on moral decision-making and AI-generated moral advice by moving beyond abstract, sacrificial dilemmas toward ecologically more valid, everyday moral scenarios. Prior work on Artificial Moral Advisors (AMAs) has predominantly focused on highly stylized hypothetical conflicts, such as trolley-type dilemmas, to examine advice effects on moral judgment and behavior (Krügel et al., 2023a, 2023b, 2025; Leib et al., 2023). While these paradigms have provided foundational insights, their limited realism constrains the transferability of findings to everyday decision-making contexts. By employing realistic, socially embedded dilemmas, the present study enhances the ecological validity of AMA research and broadens the theoretical understanding of how external advice interacts with moral reasoning in daily life.

Another novel aspect of this study is the introduction of social proximity as a moderating factor in the context of AMAs. Although social proximity has been recognized as influential in moral judgment more broadly (Haidt, 2001), its role in AI-generated moral advice remains largely unexplored. Yet, this factor is particularly interesting, as decisions involving close others often entail higher emotional stakes and heightened expectations of interpersonal sensitivity, qualities not typically associated with artificial advisors. Our findings suggest that social proximity influences behavioral responses to moral advice but does not consistently affect subjective decision experience, highlighting a nuanced interplay between relational context and moral guidance. This underscores the importance of considering relational dynamics in future research on AMAs.

Practical Implications

The present findings offer several practical insights for the responsible design and deployment of Artificial Moral Advisors (AMAs) and digital decision-support systems. The fact that individuals remain susceptible to AI-generated moral advice even (more) in realistic, socially embedded scenarios underscores the persuasive potential of such systems beyond abstract dilemmas. This highlights the need for critical reflection on the integration of AMAs into everyday decision-making contexts, particularly given users' likely limited awareness of the influence these systems may exert. Transparency about the nature and origin of advice emerges as a key consideration. Since participants were influenced by advice regardless of its advisor, users may attribute undue authority to AI advice. Ensuring clear disclosure of AI involvement and fostering critical awareness of the advisory nature of these systems are essential to prevent misplaced trust. These findings also have implications for the trustworthiness and perceived legitimacy of AI systems. While AMAs can effectively support decision-making when their advice resonates with users' internal standards, trust in AI remains fragile in contexts requiring high emotional sensitivity or moral nuance. This underscores the importance of context-aware design, transparency, and careful management of user expectations regarding the role and authority of artificial advisors. Personalization of advice holds promise for increasing user acceptance but also poses ethical challenges. Tailoring advice to individual values may risk reinforcing existing biases or discouraging moral reflection. AMAs should not merely mirror user preferences but instead foster reflective engagement and ethical deliberation. Achieving this balance requires systems that adapt to users' needs while promoting thoughtful moral reasoning.

Limitations

Despite the valuable insights gained, this study has some limitations. While we employed everyday moral dilemmas with higher ecological validity compared to abstract sacrificial scenarios, the hypothetical nature of these situations inherently limits their external validity. Participants' responses to hypothetical dilemmas may not fully capture the complexity of real-life moral decision-making. Also, our operationalization of social proximity did not evoke the level of personal involvement anticipated based on prior research (Northey et al., 2022).

We assumed that dilemmas involving socially close others would naturally elicit heightened personal engagement; however, this assumption did not hold consistently. The moral conflicts embedded in the dilemmas often positioned the participant's self-interest in direct opposition to the interests of socially close protagonists. As a result, decisions reflecting personal involvement occasionally aligned with egoistic decisions rather than altruistic ones toward close others. This misalignment suggests a conceptual oversight in our dilemma design and challenges the straightforward application of proximity-based involvement theories to moral decision-making contexts. Also, the random assignment of advice, rather than tailoring it to individual preferences or values, constrained the potential for realistic agent-user alignment. Although this approach allowed us to systematically assess the effects of advisor and framing, it reduced the ecological realism of the interaction, as real-life advice is often context-sensitive and personalized.

Future Directions

Several avenues for future research emerge from our findings. First, the choice of dilemmas warrants further refinement. While our study employed ecologically valid everyday moral scenarios, future research should consider dilemmas that evoke strong personal involvement without directly threatening participants' own interests when investigating personal engagement. Revisiting established paradigms, such as J. D. Greene et al. (2004) high-conflict moral dilemmas, may provide deeper insights into how personal involvement interacts with advice-following in morally salient contexts. Additionally, it would be valuable to investigate direct comparisons between human and AI advice within the same decision contexts, using designs that explicitly contrast the two advisors. Such studies could clarify whether certain advice characteristics uniquely enhance the credibility or influence of human or AI advisors. Building on the robust effects observed across a wide range of dilemmas of the present studies and also other research on this topic, future research should explore the nuanced influence of advice characteristics and delivery modes on moral decision-making. Beyond textual presentation, incorporating audiovisual formats or multimodal advice delivery may increase perceived authenticity, emotional resonance, or persuasive impact. Exploring whether delivery modality moderates advice-following could offer valuable insights for the design of digital

advisory systems. Finally, our results invite further investigation into the role of cognitive resources in advice-following. Specifically, future research could examine whether external advice functions as a heuristic shortcut under conditions of cognitive load or time pressure. Such studies would contribute to dual-process models of moral decision-making by clarifying whether limited deliberative capacity increases reliance on external advice.

4.6 Conclusion

Study 2 demonstrates that moral advice, independent of whether the advisor was human or artificial, can significantly influence moral decision-making, particularly when the advice aligns with internalized norms favoring altruistic decisions. The findings emphasize the central role of value congruence and social cues in advice-following, highlighting that human-provided altruistic advice exerts stronger influence in contexts involving socially distant others. Furthermore, individual dispositions such as internalization and wisdom shape both the reception of advice and subjective decision experiences. These results underscore the complexity of moral influence in human-AI interaction and suggest that the effectiveness of moral advice depends not only on advisor and framing but also on the relational and cognitive context of the decision. Study 2 thus extends prior research by illuminating how social proximity, advice characteristics, and individual traits jointly govern moral advice-following in everyday scenarios.

5 Study 3: Who Do I Follow, when my Advisors Disagree? The Effects of Convergent and Conflicting Artificial and Human Advice on Moral Decision-Making.

5.1 Abstract

Study 3 investigated how people resolve situations in which human and artificial advisors provide convergent or conflicting moral advice. In an online experiment, $N = 361$ participants completed four dilemmas (two high-conflict sacrificial and two everyday moral scenarios). On each trial, participants received two brief pieces of advice: one attributed to an AI advisor and one to a human advisor. In converging advice conditions, both advisors recommended the same option; in conflicting advice conditions, they endorsed opposing options. Decisions were recorded on a continuous scale and later dichotomized. Mixed-effects models were used to analyze decision behavior and advice-following, complemented by subjective evaluations of each advisor. Across all analyses, advice content was the primary driver of moral decision-making. When both advisors offered the same advice, participants reliably aligned with it. Effects exerted a stronger influence than in previous studies using single-source advisors. In conflicting-advice conditions, participants showed no systematic preference for either AI or human advisors; the probability of following AI or human advisors hovered near chance. Instead, advice-following was predicted by subjective appraisals: higher perceived helpfulness and pleasantness of the AI increased AI-following, whereas higher ratings for the human advisor decreased it. Demographic similarity to the human advisor selectively increased preference for human advice. These findings provide converging evidence that, in moral decision contexts, advisor identity exerts limited direct influence once advice is perceived as relevant and legitimate. Moral persuasion is driven primarily by content, perceived helpfulness, and situational fit rather than a stable bias for or against artificial advisors.

Keywords: artificial moral advice, conflicting advice, algorithm aversion & appreciation

5.2 Introduction

With the rapid rise of AI in our everyday life, research has also examined the influence of AI advisors on moral decision-making. Experimental studies indicated that when presented with moral advice from a single advisor, participants are often influenced to a similar extent, even when they are aware that the advisor is artificial (Krügel et al., 2022, 2023a, 2023b, 2025; Leib et al., 2023). AI moral advice has been found to shape moral judgments despite inconsistencies in its content or the absence of reasoned justification, with participants sometimes using it as an expedient way to resolve moral dilemmas (Krügel et al., 2023b, 2025). Similarly, the moral evaluation of others' behavior appears to depend more on the content of the advice they follow than on whether it originated from an AI or a human. These findings suggest that, at least when considered in isolation, AI and human advisors may exert comparable influence in moral domains. However, it remains unclear how people respond when both advisors provide advice in direct competition with one another. If an AI and a human advisor offer conflicting advice in the same moral decision context, will participants show a preference for one, or will the influence remain balanced?

Answering this question requires integrating insights from the broader literature on how people accept or reject AI advice. Within the research, two opposing tendencies have emerged: algorithm aversion and algorithm appreciation. Algorithm aversion refers to people's reluctance to rely on AI advice, often triggered by observing the algorithm make an error, even when it performs better than humans on average (Dietvorst et al., 2015; Jussupow et al., 2020). Algorithm appreciation, in contrast, describes the preference for AI over human advice, typically rooted in perceptions of objectivity, consistency, and accuracy (Logg et al., 2019). Both tendencies have been observed across a variety of contexts and are shaped by factors including task characteristics, perceptions of competence, transparency, error tolerance, and social presence cues (Bogert et al., 2021; Castelo et al., 2019; Hou & Jung, 2021; Longoni et al., 2019; Northey et al., 2022).

Research on algorithm aversion and appreciation reveals a complex pattern across domains. People tend to appreciate algorithms more in objective, quantitative contexts such as diagnostics or prediction, whereas aversion is more likely in subjective, creative, or moral domains (Bigman & Gray, 2018; Castelo et al., 2019). However, this pattern is not universal. Certain

subjective domains, such as dating, have been found to elicit greater appreciation for AI recommendations than human ones, suggesting that perceived suitability of AI may depend on the specific nature of the task (Logg et al., 2019).

In moral decision-making, skepticism toward AI is frequently attributed to perceived deficits in empathy, contextual sensitivity, and moral understanding (Bigman & Gray, 2018; Jussupow et al., 2024). Moral dilemmas are commonly viewed as subjective and value-laden, grounded in social expectations, interpersonal understanding, and lived experience. From this perspective, machines appear ill-equipped to grasp the experiential and emotional foundations of moral judgment. Moral evaluation is seen to require both agency and experience, capacities that artificial systems are not believed to possess. Across nine studies, participants showed robust aversion to machine-generated moral decisions in domains such as healthcare, law, and autonomous driving, even when the machine's decisions produced favorable outcomes (Bigman & Gray, 2018). Philosophical analyses converge on the same conclusion: artificial advisors do not meet the conditions required for moral responsibility, including causation, freedom, knowledge, and deliberation (Constantinescu et al., 2022). Public attitudes likewise emphasize deficits in empathy, intuition, and accountability (D. S. Kumar, 2025). Empirical work further shows that humans are attributed substantially more forward-looking responsibility and authority in ethically relevant tasks (Lima et al., 2021), and that machines are faulted less for wrongdoing, underscoring their perceived moral incompleteness (Shank et al., 2019). Scholars also warn that delegating moral reasoning to AI risks moral deskilling and the erosion of practical wisdom (Van Wynsberghe & Robbins, 2019). In total, theoretical perspectives suggest that humans should be preferred over artificial advisors in morally significant contexts.

However, recent empirical work complicates this assumption. Evidence shows that algorithm aversion in moral contexts can diminish when tasks are complex, cognitively demanding, or framed in ways that highlight algorithmic competence (Bogert et al., 2021; Hou & Jung, 2021; Jussupow et al., 2024). In some cases, AI has even been perceived as more empathetic than human counterparts (Yakar et al., 2022), challenging the view that emotional understanding necessarily advantages human advisors. Experimental research also indicates that AI moral advice can match the influence of human advice even when the algorithm's inconsistency is apparent, when its artificial nature is salient, or when the advice lacks justification (Krügel

et al., 2022, 2023a, 2023b, 2025). Leib et al. (2023) in turn, show that the moral evaluation of advice recipients is shaped more by the content of the advice than by its advisor, with only subtle differences in perceived responsibility when selfish acts follow AI rather than human advice, suggesting that advisor attributions play a limited role in shaping moral evaluations. These findings indicate that aversion to AI in moral contexts may be weaker and more context-dependent than theoretical accounts suggest.

Yet existing work has exclusively examined situations in which moral advice from humans and AI is presented in isolation, in contexts where both advisors recommend the same action or in joint decision-making. As a result, it remains unknown how people respond when human and AI advisors offer directly conflicting moral advice. Such situations reflect a more realistic decision environment, requiring individuals to reconcile competing cues and assign relative weight to each advisor. Understanding whether participants prefer one advisor over the other, or whether both exert comparable influence when placed in explicit opposition, is essential for clarifying how AI and human moral advice interact and influence in practice.

5.2.1 Aim of the Study

The present study addresses this gap by examining how participants respond to moral advice when an AI and a human advisor either provide convergent or conflicting advice. Prior research has shown that AI advice can shape moral decisions to a similar extent as human advice, even when it is inconsistent, unreasoned, or transparently artificial (Krügel et al., 2022, 2023a, 2023b, 2025; Leib et al., 2023). However, most existing studies have examined AI and human advisors in isolation, leaving open the question of how participants behave when both advisors are placed in direct competition. In situations involving conflicting advice, underlying preferences may become apparent, such as a tendency to favor human advice, greater acceptance of AI advice, or continued parity between advisors. By incorporating both convergent and conflicting advice conditions, the present study seeks to clarify the extent to which advisor identity influences advice-following in moral decision-making.

5.2.2 Hypotheses

Building on prior work (Krügel et al., 2022, 2023a, 2023b, 2025; Leib et al., 2023), recent evidence indicates that moral advice from AI can shape decision-making to a comparable extent as human advice, even when its artificial nature is salient to participants. Inconsistent or unreasoned AI recommendations have been shown to influence decisions to the same degree as human advice, suggesting that advice content often outweighs the identity of the advisor. Similarly, Y. Liu et al. (2022) demonstrated that AI and human advisors can exert equal influence in one-off moral decisions. Together, these findings suggest that algorithm aversion in moral domains may be attenuated when dilemmas lack objectively correct answers and do not entail direct personal consequences (Longoni et al., 2019; Northey et al., 2022).

The present study extends this research by placing AI and human advisors in direct competition. We examine two conditions: (a) when AI and human advisors provide convergent advice, and (b) when they provide conflicting advice within the same moral dilemma. This design allows us to test whether the observed parity in single-advisor contexts persists under direct comparison.

Based on prior findings, we derive the following hypotheses:

H1 When AI and human advisors provide the same advice, participants will be significantly more likely to follow that advice and decide accordingly. More specifically:

- ***H1a*** Altruistic advice (from both advisors) will significantly increase altruistic decisions in everyday dilemmas.
- ***H1b*** Egoistic advice (from both advisors) will significantly increase egoistic decisions in everyday dilemmas.
- ***H1c*** Utilitarian advice (from both advisors) will significantly increase utilitarian decisions in high-conflict dilemmas.
- ***H1d*** Deontological advice (from both advisors) will significantly increase deontological decisions in high-conflict decisions.

H2 In dilemmas where AI and human advisors provide conflicting advice, there will be no significant difference in the likelihood that participants follow either advisor.

5.3 Method

5.3.1 Participants

Participants were recruited through multiple channels, including social media platforms (e.g., Facebook, LinkedIn), online university platforms such as those of the University of Regensburg, SurveyCircle, university classes, and personal outreach. The final combined sample consisted of $N = 361$ participants ($M_{\text{age}} = 29.0$, $SD = 13.0$; 219 women, 135 men, 2 diverse, 5 without response). The eligibility criteria required participants to be at least 18 years old and fluent in German. Students enrolled at the University of Regensburg were eligible to receive 0.5 participant hours (*Versuchspersonenstunden*) as compensation, in line with the university's participant credit system. No financial compensation was offered to participants recruited through other channels. All participants provided informed consent before beginning the study. Ethical approval for all studies was obtained from the Ethics Committee at the University of Regensburg (approval code: 23-3326-101).

Table 5.1 displays the sociodemographic composition of the sample by gender.

5.3.2 Procedure

The study was conducted online using the SoSci Survey platform (Leiner, Dominik J., 2019) and took approximately 15 to 20 minutes to complete. Upon accessing the survey, participants first provided informed consent and confirmed their agreement with the data protection policies of the University of Regensburg. They were informed about the nature and duration of the study, their rights as participants, and the confidentiality of their responses. Participants were also advised that some scenarios might involve morally sensitive content and that there were no right or wrong answers. To familiarize participants with the task format, the survey began with a sample dilemma, the Trolley Dilemma (Foot, 1967), which was not included in the main analysis. Participants were then instructed that they would be presented with a series of moral dilemmas and asked to indicate their decisions on a continuous response scale ranging from 0 (completely Option A) to 100 (completely Option B), with each endpoint corresponding to one of the two available actions. These continuous responses were later dichotomized for analysis purposes. The main task consisted of four unique dilemmas, including two

Table 5.1: Sociodemographic Characteristics by Gender

Characteristics	**Total (N = 361)**	**Male (n = 135)**	**Female (n = 219)**	**Diverse (n = 2)**
Average age (SD)	29.0 (13.0)	33.6 (14.2)	26.0 (11.1)	21.0 (1.4)
Educational Attainment (%)				
Lower Secondary School[a]	1 (0.3%)	1 (0.7%)	0 (0.0%)	–
Intermediate Secondary School[a]	12 (3.3%)	7 (5.2%)	5 (2.3%)	–
Completed Apprenticeship	9 (2.5%)	3 (2.2%)	6 (2.7%)	–
Vocational High School	21 (5.8%)	7 (5.2%)	14 (6.4%)	–
Higher School Certificate	208 (57.6%)	57 (42.2%)	146 (66.7%)	2 (100%)
Applied Sciences University	19 (5.3%)	9 (6.7%)	9 (4.1%)	–
University Degree	91 (25.2%)	51 (37.8%)	39 (17.8%)	–
Occupation (%)				
Social Work	133 (36.8%)	26 (19.3%)	105 (48.0%)	1 (50%)
Other Activities	66 (18.3%)	23 (17.0%)	40 (18.3%)	1 (50%)
Technology	29 (8.0%)	25 (18.5%)	4 (1.8%)	–
Healthcare	55 (15.2%)	17 (12.6%)	37 (16.9%)	–
Construction & Real Estate	13 (3.6%)	8 (5.9%)	5 (2.3%)	–
Trade & Services	3 (0.8%)	2 (1.5%)	1 (0.5%)	–
Media	4 (1.1%)	3 (2.2%)	1 (0.5%)	–
Economics / Administration	23 (6.4%)	14 (10.4%)	9 (4.1%)	–
Transport & Logistics	2 (0.6%)	1 (0.7%)	1 (0.5%)	–
Natural Sciences	33 (9.1%)	16 (11.9%)	16 (7.3%)	–

[a]German school qualification levels. Gender information was missing for 5 participants and is excluded from gender-specific columns.

high conflict dilemmas and two everyday dilemmas. On each trial, participants received two pieces of advice: one attributed to an AI advisor and one to a human advisor. The content of the advice was either convergent (both advisors recommending the same option) or conflicting (each advisor recommending a different option), and the presentation order of the advisor (AI first or human first) was randomized. For each human advisor, participants were shown a photograph of the purported advisor. After reading the dilemma and the advice, participants indicated their decision using the continuous scale. Following each decision, they rated the perceived similarity and likability of the human advisor using 7 point Likert scales. A behavioral variable indicating whether participants followed the advice was computed post-hoc by comparing their decision with the direction of the advice provided. To ensure that participants were attending to the advice, a manipulation check was presented after each dilemma. Participants were asked to indicate which advisor, AI or human, gave which advice. This item was used to verify comprehension and advisor attribution on a trial-by-trial basis, providing a sensitive measure of engagement. There was no time limit imposed for reading, making decisions, or completing any part of the survey. After the main task, participants completed additional questionnaires, followed by a final debriefing. The debriefing explained the purpose of the study, clarified that all advice was simulated and standardized, and thanked participants for their contribution. This study was preregistered on the Open Science Framework: https://osf.io/ha2cj/?view_only=c9e101afe3ce4c4fa73ec8a96c970a61.

5.3.3 Measures

Dependent Variables

Participants recorded their moral decisions on a continuous scale ranging from 0 (completely Option A) to 100 (completely Option B), with each endpoint corresponding to one of the two possible actions described in the dilemma. This continuous variable was used as the primary dependent measure to capture nuanced variation in moral decision-making. For statistical modeling, this variable was also dichotomized to allow for the use of logistic regression models. Both the continuous and binary versions were used in the analyses depending on the model specification.

In addition, a second binary behavioral variable was created to indicate whether participants

followed the advice they received. In dilemmas where the AI and human advisors gave convergent advice, a response was coded as 1 if the participant's decision matched the advice, and 0 otherwise. In dilemmas with conflicting advice, two separate values were created to reflect whether the participant followed the AI or the human advisor, respectively.

Dilemmas

The dilemma materials consisted of four unique moral scenarios. Two were high conflict dilemmas adapted from J. D. Greene et al. (2004, 2008), and two were everyday moral dilemmas adapted from Singer et al. (2019). The exact dilemmas can be found in the Appendix A. Each dilemma was designed to vary along two dimensions: moral conflict type (high conflict vs. everyday) and social proximity (socially close vs. socially distant). This resulted in one dilemma for each combination of conflict level and proximity, allowing examination of how these factors jointly influence moral decision-making. All dilemmas were presented in randomized order across participants. To ensure a balanced design and to avoid emotionally polarizing content, dilemmas involving highly salient real-world topics, such as the COVID-19 pandemic or war-related conflicts, were excluded. Selection criteria prioritized dilemmas expected to produce roughly even distributions of utilitarian and deontological (or altruistic and egoistic for everyday dilemmas) responses.

Advice Stimulus

After reading each dilemma, participants received two pieces of advice: one framed as originating from an AI advisor, and the other from a human advisor. The advice consisted of a short written statement (approximately two to three sentences) recommending one of the two possible options and offering a brief justification. All advice was generated using ChatGPT (OpenAI, 2024). For each dilemma, four versions of advice were created, two supporting Option A and two supporting Option B. This was done to present two pieces of advice that differed in the wording but favored the convergent advice condition. The advisor (AI vs. human) was then randomized across participants and dilemmas, such that the same advice content could appear under either advisor label. This ensured that the tone, length, and quality of reasoning were equivalent between AI- and human-framed advice. All advice

was phrased using the subjunctive mood and framed as a suggestion rather than a directive. This linguistic framing aimed to reduce potential reactance and preserve participants' sense of autonomy in decision-making. All advice was generated in English and translated into German for presentation. Photographs were shown alongside the human advice and were randomized across trials. These images were AI-generated and taken from Nightingale and Farid (2022). They were selected and balanced to control for perceived gender (male or female) and ethnicity (Caucasian majority, South Asian, East Asian, Hispanic, and Black).

Covariates

Following each dilemma, participants rated the human advisor on two subjective dimensions using 7 point Likert scales (1 = not at all, 7 = very much): perceived similarity to the advisor and perceived likability of the advisor. In addition, participants were asked which of the two advisors they believed had more influence on their decision.

Established Instruments

After completing the main task, participants filled out three standardized questionnaires presented in a fixed order. The first was the AI Acceptance Scale (Sindermann et al., 2021), which consists of four items rated on a 7 point Likert scale. This scale measures general attitudes toward AI, with higher scores indicating greater acceptance. Next, participants completed the San Diego Wisdom Scale (SD-WISE; (Thomas et al., 2022)), which assesses individual differences in wisdom-related traits across five subdomains: decisiveness, emotional regulation, self-reflection, pro-social behavior, and acceptance of divergent perspectives. The spirituality subscale was excluded due to its lack of relevance for the research question. The SD-WISE includes 24 items, rated on a 5 point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). An example item is "I remain calm under pressure.". Finally, participants completed the Moral Identity Scale (MIS) (Aquino & Reed, 2002), which measures the degree to which morality is central to one's self-concept. It includes two subscales: Internalization (reflecting private moral identity) and Symbolization (reflecting public expression of moral traits). Each subscale contains five items, rated on a 5 point Likert scale. The scale begins with a list of nine moral trait adjectives (e.g., caring, fair, honest), followed by agreement

ratings on identity-related statements. An example item from the Internalization subscale is, "Being someone who has these characteristics is an important part of who I am.". All materials were presented in German.

5.3.4 Data Analysis

Data Preparation and Exclusion Criteria

Data preparation and analysis were conducted using RStudio (Posit Software, 2025). Data were imported from Excel files and managed using the `readxl` and `openxlsx` packages, and pre-processed with functions from the `tidyverse` ecosystem, including `dplyr` and `tidyr`, for data cleaning, recoding, and the construction of derived variables. Generalized linear mixed-effects models and linear mixed-effects models were estimated with the `lme4` package, and model-based visualizations were created using `ggplot2`. Estimated marginal means and predicted probabilities for the graphical presentation and interpretation of interaction effects were obtained using the `ggemmeans` package.

Participants who failed the manipulation check were excluded from all analyses, resulting in a final sample size of $N = 361$. Participants who did not respond to individual dilemmas were retained in the analysis, as generalized linear mixed-effects models (GLMMs) are well suited to handle missing data in repeated measures designs (Hilbert et al., 2019). This approach preserves statistical power by utilizing all available data points without requiring listwise deletion.

Experimental Design and Variable Construction

The study employed a fully within-subjects design. All participants completed the same four dilemmas. For each dilemma, participants received two pieces of advice, one from an AI and one from a human advisor. The content and order of advice were randomized across dilemmas, resulting in within-subject variation in advice congruency (convergent vs. conflicting). While the structure of the task was identical for all participants, the specific advice-advisor pairings varied by trial and individual. The key experimental predictor was the condition-specific advice received on each trial. A binary behavioral variable (advice-following) was computed post-hoc to indicate whether the participant's decision aligned with the advice. In conflicting

trials, separate indicators were used to reflect whether the AI or the human advisor was followed.

Analytic Strategy

To account for theoretical distinctions in decision-making, we divided the dataset into four analytic subsamples based on dilemma type (high conflict vs. everyday) and advice congruency (convergent vs. conflicting). This separation allowed for cleaner interpretation of results and targeted hypothesis testing. An overview of the subsample structure is provided in Table 5.2.

Table 5.2: Analytic Subsamples Based on Dilemma Type and Advice Congruency

Subsample	Description
High-conflict, convergent	Utilitarian vs. deontological dilemmas; advisors recommend the same option
High-conflict, conflicting	Utilitarian vs. deontological dilemmas; advisors recommend opposing options
Everyday, convergent	Altruistic vs. egoistic dilemmas; advisors recommend the same option
Everyday, conflicting	Altruistic vs. egoistic dilemmas; advisors recommend opposing options

Model Specification

To analyze binary outcomes (i.e., the dichotomized decision and advice-following behavior), we used generalized linear mixed-effects models (GLMMs) with a logit link function. For continuous outcomes (i.e., the original 0–100 decision scores), we used linear mixed-effects models (LMMs) with an identity link and Gaussian error distribution. In both cases, random intercepts were included for participants and dilemmas to account for individual differences and item-level variability.

The general structure of the mixed-effects models was as follows:

$$g(Y_{ij}) = \beta_0 + \beta_1(\text{Condition}_{ij}) + u_{0i} + v_{0j} + \epsilon_{ij} \tag{5.1}$$

Where:

- $g(\cdot)$ is the link function: the logit link for binary outcomes (GLMM) and the identity

link for continuous outcomes (LMM)

- Y_{ij} is the outcome for participant i on dilemma j (binary decision, advice-following behavior, or continuous decision score)
- β_0 is the fixed intercept
- β_1 represents the fixed effect of condition (e.g., advisor, congruency, or psychological predictors)
- $u_{0i} \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for participant i
- $v_{0j} \sim \mathcal{N}(0, \sigma_v^2)$ is the random intercept for dilemma j
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual error term (LMM only)

Mixed-effects modeling was selected for its robustness in analyzing hierarchical and repeated-measures data, as well as its capacity to accommodate missing responses (Al-Dajani & Czyz, 2022; Hilbert et al., 2019). All models were checked for convergence, and model performance was evaluated using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), though model comparisons were not the primary focus of this study.

In the models predicting advice-following, we included multiple predictors such as advisor similarity, likability, and individual difference measures (e.g., moral identity, wisdom, AI acceptance). Due to the number and conceptual overlap of these variables, the models initially showed signs of instability. Therefore, we assessed multicollinearity by computing the Variance Inflation Factor (VIF) for all fixed effects. Predictors with VIF values substantially exceeding recommended thresholds were examined and, if necessary, removed or restructured to ensure model stability and interpretability (James et al., 2013).

VIF diagnostics were conducted separately for high-conflict and everyday dilemma datasets. The Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to correlations with other predictors, thereby serving as a diagnostic for multicollinearity. All adjusted VIF values ($\text{GVIF}^{1/(2\times \text{Df})}$) were below conventional thresholds (e.g., < 2), indicating no serious multicollinearity concerns. For the high-conflict condition, values ranged from 1.02 (position of advice) to 1.33 (perceived pleasantness). In the everyday dilemma dataset, the highest VIF values were 1.57 for pleasantness and 1.54 for advisor

similarity. Despite conceptual overlap among variables such as likability and similarity, the overall model structure was stable and met multicollinearity assumptions.

Statistical Thresholds and Reporting

All hypotheses tests were evaluated using a conventional significance threshold of $\alpha = .05$. No corrections for multiple comparisons were applied, as all primary analyses were preregistered. For binary outcomes, results are reported as odds ratios (OR) with 95% confidence intervals and p-values. Odds ratios were computed by exponentiating the unstandardized regression coefficients ($\text{OR} = e^{\beta}$), providing an interpretable effect size metric. This allows for a straightforward interpretation of how the odds of the outcome (e.g., following advice) change with one-unit increases in the predictor variable (J. Hilbe, 2009; J. M. Hilbe, 2009). All reported ORs are interpreted relative to the appropriate reference category and accompanied by corresponding statistical significance levels, enabling meaningful comparisons between experimental conditions and psychological covariates.

Post-hoc Sensitivity Analysis

To assess the adequacy of statistical power for detecting the observed effects in our logistic regression models, a post-hoc power analysis was conducted using *G*Power* (Faul et al., 2007). Observed odds ratios (ORs) from our logistic regression analyses were used to estimate power, converted to effect size metrics compatible with G*Power's logistic regression module. For the high-conflict condition, an observed odds ratio of 8.40, an alpha level of $\alpha = .05$, and a total sample size of $n = 179$ were used as the basis for the analysis. The resulting statistical power was 1.00, indicating a very high probability of detecting the observed effect. For the everyday condition, the observed odds ratio was 5.67 with a sample size of $n = 209$ and $\alpha = .05$. Again, the analysis yielded a statistical power of 1.00.

5.4 Results

5.4.1 Summary

Across all analyses, the central factor shaping participants' moral decisions was the content of the advice, rather than its advisor.

In the convergent advice conditions, both in high-conflict and everyday moral dilemmas, participants were substantially more likely to choose the option that aligned with the advice. In high-conflict dilemmas, utilitarian advice significantly increased the likelihood of a utilitarian decision (OR $\approx$ 8.39, 95% CI [2.09, 33.76]), while in everyday dilemmas egoistic advice similarly raised the probability of egoistic decisions (OR $\approx$ 5.67, 95% CI [2.16, 14.86]). These effects were mirrored in the continuous decision scale models, confirming large mean differences between advice types.

In the conflicting advice conditions, where AI and human advisors recommended opposing options, the advisor had no measurable influence on decision outcomes in either high-conflict or everyday dilemmas. This suggests that when advice content is in direct conflict, participants do not systematically privilege one advisor over the other.

Regarding advice-following, several factors predicted the likelihood of accepting AI advice over human advice in conflicting scenarios. In high-conflict dilemmas, participants were more likely to follow AI when they rated the AI advice as more helpful and pleasant, or when the advised option was perceived as more aligned with their own moral position. Rating the human advice as more helpful, perceiving greater similarity to the human advisor, or sharing the same gender as the human advisor significantly reduced AI advice-following. In everyday dilemmas, the same pattern emerged: the more helpful participants rated the AI, the more likely they were to follow its advice, whereas higher helpfulness ratings for the human advisor decreased AI-following.

5.4.2 Descriptive Statistics

The final sample consisted of $N = 361$ participants ($n_{\text{female}} = 219$, $n_{\text{male}} = 135$, $n_{\text{diverse}} = 2$), with a mean age of $M = 29.0$ years ($SD = 13.0$). Overall, participants were generally well-educated, with the majority holding at least a higher education entrance qualification, and

many reporting a university degree. Occupations were diverse, though most were employed in social, educational, healthcare, or academic fields. Detailed distributions of education and occupational backgrounds by gender are provided in Table 5.1. The sample can be characterized as relatively young, well-educated, and socially engaged.

5.4.3 Mean Decision and Advice-Following per Dilemma

Table 5.3 presents decision rates and advice-following behavior across high-conflict and everyday dilemmas under convergent and conflicting advice conditions. Across high-conflict dilemmas, participants showed substantial variability: in the Economy dilemma, 67% chose the utilitarian option under convergent advice, while only 51% did so in the Submarine dilemma. In everyday dilemmas, altruistic decisions were less frequent in the Hamster dilemma (37%) than in the Neighbor dilemma (69%). Advice-following rates were moderate to high across conditions. Under convergent advice, around 61–68% of participants followed the provided advice regardless of dilemma type. Under conflicting advice, following rates were more balanced between AI and human advisors, ranging between 44–52% across dilemmas.

5.4.4 Main Analysis

To address our research questions systematically, we divided the main analyses into four parts, each reflecting a distinct decision context and type of statistical model. First, we examined situations in which the advice from both advisors was convergent, allowing us to assess baseline advice effects without conflicting information. Second, we analyzed conflicting advice conditions, where AI and human advice differed, to determine whether the advisor influenced decision outcomes in the face of conflict. Within each of these two decision contexts (convergent vs. conflicting), we further distinguished between high-conflict moral dilemmas and everyday moral dilemmas. This two-by-two division (convergent/conflicting × high-conflict/everyday) provides a structured framework that captures both the nature of the moral problem and the alignment of advice, enabling a direct comparison of how these factors shape advice-following.

Table 5.3: Decision Rates and Advice-Following by Dilemma Type, Advice Congruency, and Individual Dilemmas

Condition	**High Conflict Dilemmas**		**Decision Type (Everyday)**	**Everyday Dilemmas**	
	Economy	Submarine		Hamster	Neighbor
Convergent advice					
Utilitarian	67.0% (n=67)	51.1% (n=45)	Altruistic	36.8% (n=39)	69.2% (n=74)
Deontological	30.0% (n=30)	42.0% (n=37)	Egoistic	61.3% (n=65)	29.0% (n=31)
Advice follow rate	65.0% (n=65)	68.2% (n=60)		61.3% (n=65)	62.6% (n=67)
Conflicting advice					
Utilitarian	66.3% (n=120)	50.3% (n=97)	Altruistic	25.4% (n=48)	68.1% (n=128)
Deontological	30.9% (n=56)	39.9% (n=77)	Egoistic	72.0% (n=136)	27.1% (n=51)
Advice follow rate (AI)	51.9% (n=94)	46.6% (n=90)		45.5% (n=86)	48.4% (n=91)
Advice follow rate (human)	45.3% (n=82)	43.5% (n=84)		51.9% (n=98)	46.8% (n=88)

Note. Values represent percentages with counts in parentheses. Missing responses indicate selections of exactly 50 on the continuous scale (no preference). Such responses were excluded from binary decision coding, so percentages may not sum to 100%.

Convergent Advice

Empty Models We first fitted intercept-only (empty) models to examine baseline decision tendencies and to partition variance between participants and dilemmas. For both high-conflict and everyday dilemmas, the models revealed substantial random variance at both participant and dilemma levels, justifying the use of mixed-effects modeling.

Influence of Convergent Advice on Moral Decision-Making To test the effect of advisor on moral decisions, we fitted generalized linear mixed-effects models (GLMMs) for the dichotomized decision variable and linear mixed-effects models (LMMs) for the continuous decision score (0–100), with random intercepts for participants and dilemmas. In high-conflict dilemmas, participants were significantly more likely to make utilitarian decisions when receiving utilitarian advice compared to deontological advice (OR = 8.39, 95% CI [2.09, 33.76], $p = .003$). The continuous decision score confirmed this effect, with utilitarian advice lowering scores by about 24 points ($t = -5.31$). In everyday dilemmas, egoistic advice increased the likelihood of egoistic decisions relative to altruistic advice (OR = 5.67, 95% CI [2.16, 14.86], $p < .001$). Consistently, continuous decision scores were about 26 points lower under egoistic advice ($t = -5.75$). Figure 5.1 provides a visual summary of the odds ratios and confidence intervals for both dilemma types, while Table 5.4 presents the complete model estimates.

Table 5.4: Effect of AI Advice in Convergent Advice Condition on Moral Decisions

	GLMM (binary decision)			**LMM (continuous decision)**		
Dataset	β	p	*OR* [95% CI]	β	*SE*	t
High-conflict (utilitarian AI)	2.13	**.003**	8.39 [2.09, 33.76]	−23.92	4.51	−5.31
Everyday (egoistic AI)	1.74	**<.001**	5.67 [2.16, 14.86]	−25.81	4.49	−5.75

Note. Binary decision coded as utilitarian = 1 in high-conflict dilemmas and egoistic = 1 in everyday dilemmas. Continuous decision scores are coded such that higher values reflect deontological and altruistic decisions.

To aid interpretation, Figure 5.2 displays model-estimated probabilities of following the advised option for each advice type, separately for everyday and high-conflict dilemmas. The

pattern mirrors the odds-ratio results: utilitarian advice in high-conflict dilemmas and egoistic advice in everyday dilemmas were associated with substantially higher predicted probabilities of choosing in line with the advice.

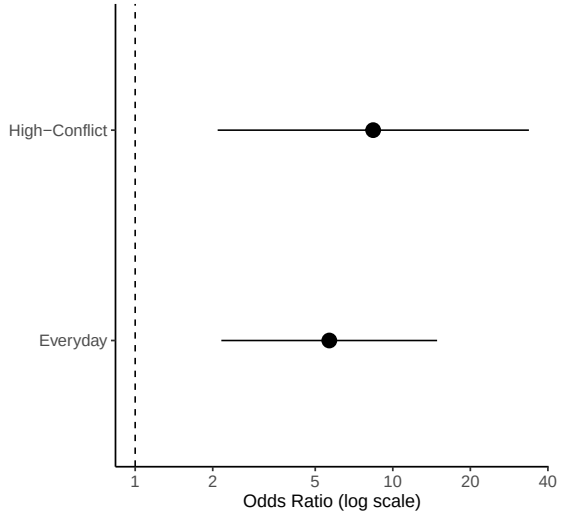


Figure 5.1: Odds ratios with 95% confidence intervals for the effect of AI advice in the equal-advice condition. Estimates above 1 indicate a higher likelihood of following utilitarian (high-conflict dilemmas) or egoistic (everyday dilemmas) advice. The vertical dashed line marks the null effect (OR = 1).

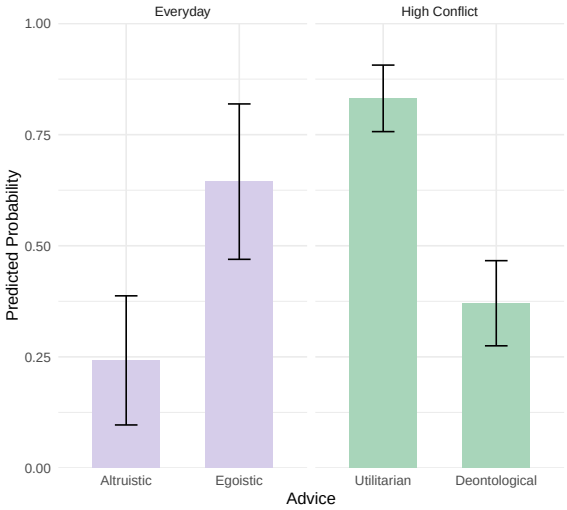


Figure 5.2: Model-estimated probability of following the advised option by advice type, shown separately for everyday and high-conflict dilemmas. Error bars represent 95% confidence intervals.

Predictors of Advice-Following in Convergent Advice Conditions We next examined individual difference predictors of advice-following in convergent advice conditions, as displayed in Table 5.5. The models included advice direction, moral identity, wisdom, AI acceptance, and the position of the advice on the screen, as well as interactions between advice and the individual difference measures (moral identity, wisdom). Random intercepts for participants and dilemmas were included. In the everyday dilemma dataset, none of the main effects of advice, moral identity, wisdom, AI acceptance, or advice position were statistically significant at the $p < .05$ threshold. Moral identity showed a positive trend ($\beta = 1.02$, $p = .075$), suggesting that participants with higher moral self-importance might have been more likely to follow advice, regardless of the advisor. No significant interactions emerged. In contrast, the high-conflict dilemma dataset revealed several significant predictors. Participants were more likely to follow utilitarian advice ($\beta = 0.80$, $p = .032$, OR = 2.22, 95% CI [1.07, 4.60]).

Higher moral identity was positively associated with advice-following ($\beta = 1.38$, $p = .046$, OR = 3.99, 95% CI [1.03, 15.45]), whereas higher wisdom scores predicted lower advice-following ($\beta = -2.69$, $p = .001$, OR = 0.07, 95% CI [0.02, 0.34]). Positive attitudes toward AI also predicted higher advice-following ($\beta = 0.37$, $p = .031$, OR = 1.44, 95% CI [1.04, 2.00]). Interactions between advisor and moral identity or wisdom were not significant in either dataset. These findings suggest that advisor and individual difference measures play a more pronounced role in high-conflict moral contexts than in everyday dilemmas.

Table 5.5: Predictors of Advice-Following in Convergent-Advice Condition

Predictor	**Everyday dilemmas**			**High-conflict dilemmas**		
	β	p	OR [95% CI]	β	p	OR [95% CI]
Intercept	1.01	.215	2.76 [0.55, 13.86]	−1.05	.230	0.35 [0.07, 1.73]
AI advice (egoistic/utilitarian)	−0.34	.259	0.71 [0.39, 1.28]	0.80	**.032**	2.22 [1.07, 4.60]
Moral identity	1.02	.075	2.77 [0.91, 8.44]	1.38	**.046**	3.99 [1.03, 15.45]
Wisdom	−0.01	.982	0.99 [0.29, 3.41]	−2.69	**.001**	0.07 [0.02, 0.34]
AI acceptance	−0.14	.348	0.87 [0.64, 1.18]	0.37	**.031**	1.44 [1.04, 2.00]
Advice position	0.19	.540	1.21 [0.66, 2.21]	0.02	.960	1.02 [0.50, 2.07]
AI advice × moral identity	−1.02	.160	0.36 [0.09, 1.45]	−0.82	.352	0.44 [0.08, 2.42]
AI advice × wisdom	0.04	.967	1.04 [0.20, 5.39]	1.54	.177	4.66 [0.50, 43.49]

Note. β = unstandardized log-odds coefficient; OR = odds ratio; CI = confidence interval. Bold p values indicate significance at $p < .05$. Moral identity and wisdom are grand-mean centered.

Conflicting Advice

Empty Models for Conflicting Advice Conditions We first estimated empty (null) models for the conflicting-advice datasets to partition variance between participants and dilemmas. Intercept-only models showed that in everyday dilemmas, baseline decisions did not differ significantly between participants, with most variance attributable to dilemmas. In high-conflict dilemmas, baseline decisions were slightly biased toward the reference option, with variance arising at both the participant and dilemma levels. These models provide the baseline

against which the effects of advisor and other predictors were tested.

Effect of Advisor on Decisions in Conflicting Advice Conditions We examined whether the advisor influenced decision-making when advice was conflicting. Table 5.6 shows the results from the GLMM and LLM for conflicting advice.

For high-conflict dilemmas, the generalized linear mixed-effects model revealed no significant effect of advisor on binary decisions, $\beta = 0.13$, $SE = 0.24$, $z = 0.51$, $p = .608$. The intercept approached significance ($\beta = 0.47$, $SE = 0.24$, $p = .055$), indicating a slightly higher (non significant) baseline probability of choosing the reference decision. The linear mixed-effects model with the continuous decision measure confirmed this null effect, showing no significant difference between advisors, $\beta = -1.80$, $SE = 2.93$, $t = -0.61$. For everyday dilemmas, the binomial model also showed no significant influence of advisor, $\beta = 0.06$, $SE = 0.23$, $z = 0.24$, $p = .814$, and the intercept was non-significant ($\beta = -0.09$, $SE = 0.71$, $p = .901$).

Consistently, the continuous decision model indicated no meaningful difference, $\beta = 0.25$, $SE = 3.37$, $t = 0.08$. Across both dilemma types and both measurement scales (binary and continuous), the advisor did not significantly affect participants' decisions when the advice was conflicting.

Table 5.6: Effect of AI Advice in Conflicting Advice Condition on Moral Decisions

	GLMM (binary decision)			**LMM (continuous decision)**		
Dataset	β	p	OR [95% CI]	β	SE	t
High-conflict (utilitarian AI)	−0.13	.608	0.88 [0.55, 1.42]	−1.80	2.93	−0.61
Everyday (egoistic AI)	0.06	.814	1.06 [0.67, 1.67]	0.25	3.37	0.08

Note. Binary outcomes reflect the probability of deontological choices (high-conflict) and altruistic choices (everyday). Continuous scores (0–100) increase with deontological/ altruistic decisions. No significant effects of AI advice were observed; everyday models showed boundary fits.

As shown in Figure 5.4, participants followed AI- and human-provided advice at roughly equal rates when the two advisors disagreed. Across both high-conflict and everyday dilemmas, the predicted probabilities of following the AI or the human advisor hovered near chance level (approximately .50), with no meaningful differences between utilitarian versus deontological

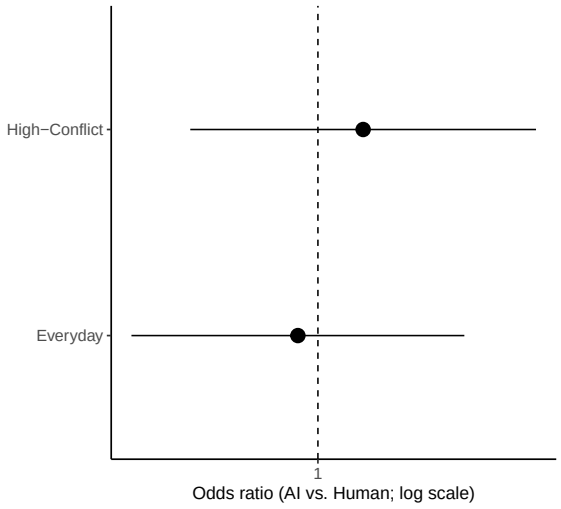


Figure 5.3: Odds ratios (OR) with 95% confidence intervals for the effect of AI advice (vs. human advice) in conflicting-advice conditions, separately for high-conflict and everyday dilemmas. The vertical dashed line indicates the null effect (OR = 1).

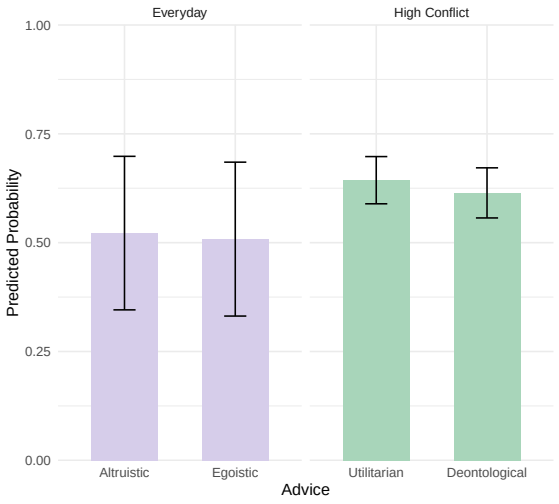


Figure 5.4: Predicted probability of following egoistic/ utilitarian advice in conflicting-advice conditions, separately for everyday and high-conflict dilemmas. Error bars represent 95% confidence intervals.

advice in high-conflict dilemmas or between egoistic versus altruistic advice in everyday dilemmas. This descriptive pattern is consistent with the nonsignificant fixed effects reported in Table 5.6. The corresponding odds ratios are displayed in Figure 5.3. For both dilemma types, the odds ratios clustered around 1 and their 95% confidence intervals crossed the null line, indicating that participants were equally likely to follow either the AI or the human advisor when their advice conflicted.

Prediction of Advisor Preference in Conflicting Advice Conditions Next, we examined which factors predicted whether participants followed the AI's advice (vs. the human's advice) when both advisors provided conflicting advice. Complete model results across all dilemmas are presented in Table 5.7.

In high-conflict dilemmas, several factors significantly influenced AI advice-following. Participants were less likely to follow AI advice when the advisor was of the same sex (OR = 0.52, 95% CI [0.29, 0.93], $p = .029$) or perceived as more similar to one self (OR = 0.68, 95% CI [0.51, 0.92], $p = .013$). As expected, perceived helpfulness strongly predicted decisions: participants followed AI advice more when the AI was rated as more helpful (OR = 2.18, 95% CI [1.61, 2.95], $p < .001$), and less when the human was rated as more helpful (OR = 0.46,

95% CI [0.34, 0.64], $p < .001$). Perceptions of influence also mattered: perceiving the AI as more influential than the human increased the likelihood of following AI advice (OR = 5.72, 95% CI [1.26, 26.00], $p = .024$). Finally, higher AI acceptance was associated with greater AI advice following (OR = 1.35, 95% CI [1.03, 1.77], $p = .032$). A similar but weaker pattern emerged in everyday dilemmas. AI helpfulness again promoted following AI advice (OR = 2.19, 95% CI [1.54, 3.12], $p < .001$), while human helpfulness reduced it (OR = 0.53, 95% CI [0.37, 0.77], $p < .001$). Participants were less likely to follow AI advice when the human was perceived as more influential (OR = 0.23, 95% CI [0.05, 0.95], $p = .043$). Perceived similarity showed a marginally negative association (OR = 0.75, 95% CI [0.54, 1.04], $p = .087$). Across both dilemma types, perceived helpfulness of the AI and human advisors emerged as the strongest predictors of advisor preference. Additional effects, such as similarity, influence ratings, and AI acceptance were evident in high-conflict dilemmas, whereas everyday dilemmas showed a more restricted set of predictors.

Table 5.7: Predictors of Following AI Advice (vs. Human Advice) in Conflicting Advice Conditions

Predictor	High-conflict dilemmas			Everyday dilemmas		
	β	p	OR [95% CI]	β	p	OR [95% CI]
Intercept	−0.93	.300	0.40 [0.07, 2.28]	1.68	.117	5.37 [0.66, 44.08]
Same sex	−0.65	**.029**	0.52 [0.29, 0.93]	−0.18	.579	0.84 [0.45, 1.56]
Social proximity (distant)	0.10	.720	1.10 [0.64, 1.89]	−0.27	.395	0.76 [0.41, 1.42]
Pleasantness	0.36	**.021**	1.44 [1.06, 1.96]	−0.24	.193	0.79 [0.55, 1.13]
Similarity	−0.38	**.013**	0.68 [0.51, 0.92]	−0.29	.087	0.75 [0.54, 1.04]
Helpfulness (AI)	0.78	**<.001**	2.18 [1.61, 2.95]	0.78	**<.001**	2.19 [1.54, 3.12]
Helpfulness (human)	−0.77	**<.001**	0.46 [0.34, 0.64]	−0.63	**<.001**	0.53 [0.37, 0.77]
Influence: equal impact	0.50	.405	1.65 [0.51, 5.42]	−0.69	.356	0.50 [0.12, 2.18]
Influence: more AI	1.74	**.024**	5.72 [1.26, 26.00]	0.36	.655	1.43 [0.30, 6.96]
Influence: more human	0.56	.385	1.76 [0.49, 6.29]	−1.49	**.043**	0.23 [0.05, 0.95]
Influence: no impact	0.96	.107	2.62 [0.81, 8.41]	−0.46	.468	0.63 [0.18, 2.20]
AI acceptance	0.30	**.032**	1.35 [1.03, 1.77]	−0.10	.498	0.90 [0.67, 1.22]
Position of advice	0.39	.160	1.48 [0.86, 2.56]	0.31	.330	1.37 [0.73, 2.58]

Note. β = unstandardized log-odds coefficient; OR = odds ratio; CI = confidence interval. Bold p values indicate statistical significance at $p < .05$.

5.5 Discussion

5.5.1 Summary of Key Findings

The present study examined the relative influence of advice content and advisor on moral decision-making across two types of dilemmas (high-conflict and everyday moral dilemmas) and under two advice conditions: convergent and conflicting advice from AI and human advisors.

Across all analyses, the content of the advice emerged as the primary determinant of the moral decision, with the advisor exerting no measurable influence on decision outcomes. In

convergent-advice conditions, participants reliably selected the option aligned with the advice provided by both AI and human advisor. Specifically, in high-conflict dilemmas, receiving utilitarian advice increased the likelihood of making a utilitarian decision by a factor of approximately 8.39 ($p < .01$) relative to the baseline. In everyday dilemmas, receiving egoistic advice increased the likelihood of making an egoistic decision by a factor of approximately 5.67 ($p < .001$). These patterns were corroborated by continuous decision-scale models, which revealed large mean differences between advice types.

In conflicting-advice conditions, where AI and human advisors recommended opposing courses of action, participants showed no systematic preference for either advisor, eventhough human advisors were visually represented with photographs to activate social and demographic cues. Instead, advice-following was driven by subjective evaluations: higher perceived helpfulness and pleasantness of the AI advisor increased the probability of following its advice, whereas higher ratings for the human advisor decreased it. Moreover, demographic congruence had a targeted effect: sharing the same gender as the human advisor or perceiving the human advisor more similar to oneself significantly increased the likelihood of favoring human over AI advice.

5.5.2 Main Effects of Advice Convergence

Our findings demonstrate that convergent advice exerted a strong and highly significant influence on moral decision-making. When both advisors recommended the same course of action, participants were substantially more likely to follow that advice, replicating the pattern observed in my previous studies. These new results extend earlier findings by showing that advice from two converging advisors is even more persuasive than advice from a single advisor. We propose that this amplification is driven by the accumulation of persuasive cues: two advisors advocating the same decision not only double the amount of argumentation but also provide stronger social proof. From a dual-process perspective, convergent signals likely reduce cognitive conflict and lower the perceived need for further deliberation, thereby making the advised option more compelling.

5.5.3 Absence of Main Effects in Conflicting Advice

In contrast, conflicting advice did not affect the overall decision pattern. Participants showed no systematic preference for either AI or human advisor when their advice diverged. This aligns with prior findings that advice use is guided less by the advisor's identity and more by the expedience of the advice itself (Krügel et al., 2022, 2023a, 2023b; Yeomans et al., 2019). Rather than reflecting stable advisor preferences, advice selection appears to operate as a pragmatic heuristic for effort reduction (Shah & Oppenheimer, 2008). Notably, perceived helpfulness emerged as a consistent predictor of advice-following, regardless of whether the advice came from a human or AI. This subjective evaluation mirrors earlier results showing that advice perceived as useful is more likely to be followed (Logg et al., 2019). However, our data also suggest the possibility of reverse causality: participants may have rated the advisor as "helpful" primarily because their advice aligned with the participants' pre-existing preferences. This pattern is consistent with research on confirmation bias and post-hoc rationalization in intuitive moral reasoning (Haidt, 2001; Mercier, 2011), and complicates the interpretation of perceived helpfulness as an independent causal factor. Crucially, the absence of a main effect of advisor type underscores that AI is no longer viewed as categorically less trustworthy or authoritative than human advisors, especially when artificial advice possibly reinforces the own judgment. While some studies have reported lingering distrust toward AI advice in high-stakes contexts (Bigman & Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2015; Jussupow et al., 2020), others show that individuals are increasingly willing to follow AI advice (Bogert et al., 2021; Logg et al., 2019), especially when framed as competent or empathetic (Bigman & Gray, 2018). Our findings contribute to this evolving picture by showing that, in moral decision contexts, advisor identity alone does not predict advice acceptance.

5.5.4 Role of AI Acceptance and Trust

Individual differences in AI acceptance further support this interpretation: participants who reported greater openness toward AI were more likely to follow AI advice, even in conflicting situations. This suggests that attitudinal trust modulates advice-following, though it does not create a systematic preference for AI at the group level. These results are consistent with prior work showing that individuals with more favorable views of AI systems are generally

more inclined to accept their advice (Logg et al., 2019).

5.5.5 Similarity-Based Moderators

A particularly informative pattern emerged in the conflicting-advice condition: participants were less inclined to follow AI advice when the human advisor shared salient social similarities with them, such as the same gender or high perceived interpersonal similarity. This aligns with research on similarity–attraction effects and in-group favoritism, which consistently show that people are more receptive to input from socially similar others (Byrne, 1997; Goette et al., 2006). These cues of social proximity appear to partially offset the otherwise neutral stance toward AI in moral advice contexts. In practical terms, a human advisor perceived as socially similar retains a persuasive advantage over an equally credible AI advisor, underscoring the persistent role of group-based biases in advice-following.

5.5.6 Moral Identity and Wisdom

Beyond situational factors, individual differences in moral identity predicted advice-following in convergent-advice contexts. For participants who strongly value seeing themselves as moral individuals, accepting the joint advice of both advisors may have served as a way to affirm their alignment with socially endorsed moral norms. Rejecting convergent moral advice, by contrast, may have posed a threat to their moral self-concept (Aquino & Reed, 2002; Hardy & Carlo, 2005). Unexpectedly, trait wisdom was negatively associated with advice-following in the convergent-advice condition, yet reversing pattern were found in earlier studies. One possibility is that the high-conflict dilemmas used in this study introduced contextual features that altered how wise individuals approached external advice. Prior research suggests that wisdom is associated with autonomy, perspective-taking, and resistance to conformity when moral complexity is high (Grossmann et al., 2010; Staudinger et al., 1992). It is therefore conceivable that individuals high in wisdom viewed converging advice as a form of external pressure, prompting greater scrutiny or independent reasoning rather than compliance.

5.5.7 Implications

The present findings have significant implications for our understanding of moral persuasion and for the design and governance of AI systems that provide moral advice. Theoretically, the study extends the existing research by adding directly conflicting human and artificial advice. These results challenge the notion that humans possess a unique capacity to influence moral decision-making. Across a range of classic and novel dilemmas, AI-generated advice proved as persuasive as human advice, even in high-conflict and high-stakes scenarios. This suggests that advice-following may be less dependent on the advisor's perceived humanity or authority, and more on the framing and clarity of the content itself.

From a practical perspective, this equivalence highlights the persuasive potential of AI in ethically sensitive contexts such as healthcare, education, or legal decision-making. Since users may respond to AI moral advice as readily as to human advice, designers and policymakers must treat AI systems not just as neutral information tools but as normative agents capable of shaping user behavior and values (Constantinescu et al., 2022; Formosa & Ryan, 2021). The central role of content in driving moral decisions implies that small variations in phrasing can produce significant behavioral effects. As such, the framing of AI advice must be subjected to careful ethical scrutiny, with mechanisms in place to audit, monitor, and refine advice outputs over time. This is especially important given AI's scalability and consistency, which amplify its capacity to shape public norms through cumulative exposure.

Ethically, these findings raise concerns about user autonomy and susceptibility to unexamined influence. If users fail to distinguish between human and artificial advice, or accept both uncritically there is a risk of moral delegation, in which individuals outsource ethical reflection to AI systems. This may foster ethical passivity and reduce engagement with morally complex decisions. Further concerns include the reinforcement of dominant moral norms via adaptive learning systems, which may marginalize minority perspectives, and the use of personalization techniques that exploit demographic congruence effects without users' awareness. Together, these risks point to the need for strong governance frameworks that go beyond technical accuracy to include normative alignment, cultural sensitivity, and continuous ethical oversight (Agbese et al., 2021; Binns, 2018; Crawford, 2021; Hossain & Habib, 2024).

Methodological Reflections and Limitations

While the present study offers novel insights, several methodological limitations warrant careful consideration. A central limitation concerns the interpretation of perceived helpfulness ratings. There is a risk that participants may have rated as "helpful" the advisor whose advice aligned with their final decision. This introduces a risk of post-hoc justification rather than reflecting a genuine evaluation of the advisor. In other words, self-reported helpfulness may be endogenous to participants' final choice, thereby obscuring the causal direction between perceived helpfulness and advice-following. Another limitation is the reliance on self-report measures for psychological constructs like wisdom. These measures are inherently prone to reporting biases, including subjective distortion, limited introspective access, and consistency pressures. Such biases can affect the validity of conclusions drawn from these constructs (Donaldson & Grant-Vallone, 2002). A separate concern relates to social desirability bias, which may particularly influence responses in moral dilemma tasks. In such contexts, participants might provide answers they perceive to be more socially acceptable, rather than reflecting their authentic moral intuitions or decision-making processes. Also, the sample consisted primarily of psychologically interested or educated individuals recruited through online platforms and the university. In addition, psychology students received credit for participating. This introduces a potential selection bias. The sample may differ systematically from the general population in terms of cognitive style, openness to advice, or familiarity with classic psychological paradigms. As a result, the generalizability of our findings to broader, less psychology-literate populations is limited. Moreover, online studies inherently lack full experimental control. Participants completed the study in varying environments (e.g., at home, in transit), which may have introduced uncontrolled noise into response patterns. This includes variations in attention span, distractions, and screen size, all of which can affect task engagement and data quality.

5.5.8 Future Research

Building on the current findings, future research should investigate how moral advice-taking unfolds under more ecologically valid and cognitively constrained conditions. Most notably, the roles of time pressure and cognitive load require systematic exploration. According to

the Elaboration Likelihood Model (ELM), individuals under cognitive strain are less likely to engage in central (i.e., content-driven) processing, and more likely to rely on peripheral cues, such as advisor identity or perceived authority. We conceptualize advice-following as a cognitive shortcut that helps minimize mental effort during decision-making (Shah & Oppenheimer, 2008). These theoretical perspectives jointly suggest that in high-load or time-limited situations users may rely more heavily on advice, potentially amplifying both beneficial and manipulative effects of AI advice.

A second important avenue involves the timing of subjective evaluations. In the present design, advice was rated only after a moral decision had been made, which introduces the risk of post-hoc rationalization. Future studies should experimentally manipulate when advice evaluations are solicited, in order to disentangle whether perceived helpfulness is a driver or a retrospective justification of the decision.

In addition, future work should examine how the modality of moral advice delivery (textual, audio, or video-based) influences its reception. Moral advice is often embedded in socially rich, dynamic interactions rather than static text. Video-based or conversational agents may evoke stronger engagement, social resonance, or trust than written formats. Research comparing modalities could reveal whether richer communication channels enhance or mitigate persuasive effects, and whether this varies by advisor or moral frame. Finally, questions around transparency and trust remain central. While transparency is often assumed to enhance trust, prior findings suggest that in moral contexts, elaboration or justification may actually reduce the persuasive impact of advice. Future studies should explore how users perceive and respond to varying degrees of algorithmic transparency, and whether such disclosures interact with advice modality, content framing, or individual differences in moral identity, autonomy, and AI literacy.

5.6 Conclusion

The present study provides robust evidence that in moral decision-making, the decisive factor is the content of advice rather than its advisor. Across both high-conflict and everyday dilemmas, participants reliably aligned their decisions with the direction of the advice, whether it originated from a human or an AI. In direct, head-to-head comparisons where AI and

human advisors provided conflicting advice, no systematic preference for the human advisor emerged. These results challenge the assumption that humans inherently resist AI advice in morally relevant contexts and suggest that advice functions primarily as a heuristic, enabling individuals to reduce cognitive effort in complex decisions. While demographic similarity, such as sharing the same gender with a human advisor, and subjective evaluations like perceived helpfulness and similarity, influenced advice-following, these effects were secondary to the persuasive weight of the advice content. The absence of a consistent human advantage raises important theoretical, practical, and ethical considerations. If humans do not reliably act as a corrective to AI influence, the design, framing, and value alignment of AI moral advice systems gain heightened significance. As AI systems increasingly enter decision-making domains, understanding the mechanisms of advice acceptance and especially safeguarding moral autonomy becomes essential. Future research should explore cultural variability, ecological validity, and the long-term effects of repeated exposure to AI moral advice, ensuring that these systems support rather than erode human moral agency.

6 Study 4: Can AI Change my Mind if I Have Already Decided? The Effect of Artificial Advice on Retrospective Decision-Change.

6.1 Abstract

Study 4 examined whether moral advice can retrospectively alter already-formed moral decisions and whether such post-decisional influence differs between human and artificial advisors. In an online experiment, $N = 356$ participants completed four dilemmas. For each dilemma, participants first made a binary decision, then received brief advice framed as either an AI or a human advisor that either supported or opposed their initial decision, and subsequently indicated whether they would make the same decision again. Mixed-effects models were used to predict decision consistency and changes over time. Overall, initial moral decisions were relatively stable across the whole sample, yet advice that opposed the initial decision significantly reduced decision consistency. This destabilizing effect was observed for both everyday and sacrificial dilemmas but was asymmetrical with respect to normative content: advice promoting altruistic or deontological options more readily corrected prior egoistic or utilitarian decisions than vice versa. By contrast, advice encouraging norm-deviant options was often dismissed, especially when it challenged deontological stances in sacrificial dilemmas. Advisor type did not systematically moderate these effects, whereas perceived helpfulness emerged as the most consistent predictor of retrospective advice-following. The findings show that post-decisional moral advice can shift previously formed decisions, particularly when it offers norm-congruent, socially sanctioned alternatives. Artificial and human advisors exert comparable influence in this retrospective context, reinforcing the conclusion that the persuasive power of moral advice depends more on its framing and perceived usefulness than on whether it is attributed to an AI or a human.

Keywords: retrospective moral decision-change, artificial advice, cognitive dissonance

6 Study 4: Can AI Change my Mind if I Have Already Decided? The Effect of Artificial Advice on Retrospective Decision-Change.

6.2 Introduction

6.2.1 Intuitive Moral Decisions and Post-Decisional Change

Moral decisions typically arise rapidly and intuitively, grounded in affective reactions rather than extended deliberation. According to the social intuitionist model, evaluative responses to moral scenarios are driven by evolved sensitivities to harm, fairness, and loyalty, whereas explicit reasoning mainly serves to justify intuitive verdicts after the fact (Haidt, 2001; Haidt & Joseph, 2004). Dual-process accounts converge on this view, suggesting that deontological responses are tied to emotion-laden, automatic processes, whereas utilitarian choices depend more heavily on controlled, prefrontally mediated cognition (J. D. Greene et al., 2004, 2008).

At first glance, such models imply that once an intuitive moral decision is formed, it should be comparatively stable. Yet moral decisions do change, particularly when individuals encounter new information. In these post-decisional contexts, people face a conflict between preserving coherence with their prior choice and integrating external input that challenges it. Cognitive dissonance theory conceptualizes this as tension between an existing commitment and dissonant evidence, motivating people either to revise their decision or to defend it in order to restore psychological equilibrium (Festinger, 1962). In moral domains, this tension is amplified by the need to uphold a positive moral identity (Aquino & Reed, 2002; Monin & Jordan, 2009). Individuals are especially likely to reject feedback that threatens their perceived moral adequacy, thereby stabilizing existing judgments but limiting opportunities for moral change (Ditto et al., 2009).

Research on anchoring and advice-taking highlights similar dynamics. Initial decisions exert a strong anchoring influence on subsequent evaluations (Tversky & Kahneman, 1974), and people systematically underweight advice that contradicts their prior choice, particularly once a decision has been made (Schultze et al., 2017). Meta-analytic evidence shows that supportive advice is perceived as more competent and trustworthy, whereas opposing advice is more often dismissed (Bailey et al., 2023). Moreover, advice presented after an initial decision receives less weight than advice presented beforehand, consistent with the idea that post-decisional input confronts a motivationally fortified judgment (Rebholz & Hütter, 2022). Still, dual-process and coherence-based approaches both imply that revision remains possible:

controlled deliberation can override intuitive responses under conditions of conflict (J. D. Greene et al., 2008), and belief systems may reorganize dynamically to minimize internal inconsistency (Holyoak, 1999; D. Simon et al., 2004).

Social framing is crucial for understanding when such reorganization occurs. Prescriptive moral messages that emphasize altruism or fairness tend to elicit affirmation and approach motivation, whereas proscriptive or norm-deviant cues (e.g., advice advocating self-interest or harm) can trigger guilt, defensiveness, or reactance (Janoff-Bulman et al., 2009). Conformity pressures further encourage alignment with perceived social norms (Cialdini & Goldstein, 2004), and impression-management motives lead people to favor judgments that preserve a valued moral self-image, especially under conditions of accountability (Tetlock, 2002). Within this framework, advice that reinforces culturally sanctioned norms, such as altruistic responses in everyday dilemmas or deontological prohibitions in sacrificial dilemmas, may function as a legitimate corrective that facilitates moral realignment without threatening moral identity.

6.2.2 Artificial and Human Advisors in Post-Decisional Moral Contexts

Artificial Intelligence has increasingly assumed advisory roles in many domains with also playing an emerging role in moral decision contexts (Cervantes et al., 2020; Constantinescu et al., 2022; Fabre et al., 2024; Formosa & Ryan, 2021). Public responses to AI advisors oscillate between algorithm aversion, the reluctance to rely on algorithms even when they outperform humans, and algorithm appreciation, the tendency to trust them when they are perceived as objective or expert (Dietvorst et al., 2015; Logg et al., 2019). Acceptance depends less on objective accuracy than on perceived competence, fairness, transparency, and process comprehensibility (Ng et al., 2021; Yeomans et al., 2019). Explanations that enhance subjective understanding substantially reduce aversion, suggesting that trust hinges on cognitive fluency rather than performance alone.

Conceptual work on machine morality distinguishes between AI as a moral agent, as a moral patient, and as a moral proxy mediating human interactions (Bonnefon et al., 2020). As moral proxies, AI systems may provide guidance in ethically charged situations and thus act as Artificial Moral Advisors. However, the notion of stable, principle-driven AMAs clashes with empirical findings that human moral cognition is flexible, context-sensitive, and dynamically

constructed (Y. Liu et al., 2022). Moral advice acceptance itself constitutes a moral judgment: individuals selectively adopt advice consistent with their pre-existing values while rejecting incompatible input, potentially generating moral echo chambers and risking moral deskilling when external guidance is overused.

Recent empirical work shows that AI-generated advice can influence moral judgments and behavior to a degree comparable to human advice. AI and human advisors have been found to exert similar sway over moral choices, even when the AI is portrayed as untrustworthy or inconsistent (Krügel et al., 2022, 2023b, 2025). Likewise, AI advice encouraging dishonest or unethical behavior increases rule violations and transparency about the artificial source does not reliably attenuate these effects (Leib et al., 2021, 2023). Meta-analytic evidence suggests that such patterns are best understood in terms of advice "fit": persuasive strength depends on the alignment between advisor characteristics, task demands, and decision-maker expectations rather than on human versus algorithmic origin per se (Bailey et al., 2023; Baines et al., 2024). To date, research on AI moral persuasion has largely focused on situations in which advice is presented prior to the formation of a judgment. Consequently, considerably less is known about whether artificial advisors can retrospectively shape moral decisions that have already been made, or whether such post-decisional influence differs from that exerted by human advisors.

Yet, in everyday contexts, individuals rarely encounter moral advice in a state of neutrality. Rather, they typically generate an intuitive evaluation early in the process, shaped by affective reactions, habitual reasoning patterns, and prevailing social norms (Haidt, 2001; Haidt & Joseph, 2004). Any subsequent advice must therefore contend with established anchors as well as the motivational pressures associated with post-decisional dissonance, including the drive to justify one's initial choice and to maintain a coherent and morally adequate self-concept (Festinger, 1962). Moral judgments may thus originate in affective intuition but are subsequently stabilized through self-consistency motives. External advice can challenge these initial intuitions, with its acceptance depending on factors such as perceived helpfulness, framing, and advisor credibility. As AI systems increasingly enter domains involving moral evaluation, understanding how individuals integrate, resist, or reinterpret AI input becomes critical for both psychological theory and ethical design. The present study addresses these

issues by examining whether and how moral advice can alter previously formed decisions, and whether such post-decisional influence depends on the advisor being human or artificial.

6.2.3 Aim of the Study

The present study investigates whether and under which conditions moral advice can retrospectively alter already formed moral decisions, and whether such post-decisional influence differs between human and artificial advisors. Building on social–intuitionist and dual-process models, we aim to create situations in which individuals first make an initial intuitive moral choice and only afterwards receive moral advice that either supports or opposes their initial decision. The study systematically varies dilemma type (everyday vs. sacrificial), advice content (altruistic vs. egoistic; deontological vs. utilitarian), advice congruence (supportive vs. opposing), and advisor (human vs. AI) in order to examine (a) the overall stability of initial moral decisions under post-decisional advice, (b) potential asymmetries between norm-congruent (altruistic, deontological) and norm-deviant (egoistic, utilitarian) advice, and (c) the relative persuasive strength of AI compared to human moral advice.

Hypotheses

Based on the theoretical framework and prior empirical findings, we formulated the following hypotheses:

***H1*: Retrospective Influence of Advice.** Across dilemma types, post-decisional moral advice will exert a significant retrospective influence on moral decision-making, such that participants' post-advice decisions systematically shift in the direction of the received advice.

***H1a*: Human Advice.** Human moral advice will exert a significant retrospective influence on moral decisions, leading to systematic shifts in post-advice decisions relative to participants' initial choices.

***H1b*: Artificial Advice.** Artificial (AI-generated) moral advice will likewise exert a significant retrospective influence on moral decisions, leading to systematic shifts in post-advice decisions relative to participants' initial choices.

***H1c*: Advisor equivalence.** The overall magnitude of retrospective influence will not differ reliably between human and AI advisors; that is, human and artificial moral advice will be comparable in their capacity to change post-decisional moral decisions.

***H1d*: Normative Content Asymmetry.** Based on our previous experiments, we expect that advice that aligns with culturally sanctioned moral norms (altruistic in everyday dilemmas; deontological in sacrificial dilemmas) will exert stronger retrospective influence on moral decisions than norm-deviant advice (egoistic or utilitarian), particularly when the advice opposes the initial decision.

6.3 Method

6.3.1 Participants

We recruited $N = 358$ participants from the University of Regensburg. The study was conducted online in German. Sociodemographic information (age, gender, education, occupation) was collected at the end of the survey. Table 6.1 reports the sociodemographic characteristics by gender. The mean age of the sample was 29.0 years ($SD = 13.0$). Participants received course credit ("Versuchspersonenstunden") and no monetary compensation. Inclusion criteria required participants to be at least 18 years old and fluent in German. Demographic variables (age, gender, education, occupation) and individual-difference measures of AI aversion, wisdom, and moral identity were collected. All participants provided informed consent prior to participation.

6.3.2 Procedure

The study was administered online using the SoSci Survey platform (Leiner, Dominik J., 2019) and took approximately 15 minutes to complete. After providing informed consent, participants were informed about the general study purpose, the confidential handling of their data, and their right to withdraw at any time. They were also advised that some scenarios might involve morally sensitive content and that there were no objectively correct answers.

Participants then completed four trials in randomized order. Each trial followed the same sequence: (1) reading the dilemma, (2) making an initial binary moral decision (Option A vs.

Table 6.1: Sociodemographic Characteristics by Gender

Characteristics	**Total (N = 356)**	**Male (n = 135)**	**Female (n = 219)**	**Diverse (n = 2)**
Average age (SD)	29.0 (13.0)	33.6 (14.2)	26.0 (11.1)	21.0 (1.41)
Educational attainment (%)				
No degree	0 (0.0%)	–	–	–
Lower secondary school[a]	1 (0.3%)	1 (0.7%)	–	–
Intermediate secondary school[a]	12 (3.4%)	7 (5.2%)	5 (2.3%)	–
Completed apprenticeship	9 (2.5%)	3 (2.2%)	6 (2.7%)	–
Vocational high school	21 (5.9%)	7 (5.2%)	14 (6.4%)	–
Higher school certificate (Abitur)	205 (57.6%)	57 (42.2%)	146 (66.7%)	2 (100.0%)
Applied sciences degree	18 (5.1%)	9 (6.7%)	9 (4.1%)	–
University degree	90 (25.3%)	51 (37.8%)	39 (17.8%)	–
Occupation (%)				
Social work and education	132 (37.1%)	26 (19.3%)	105 (48.0%)	1 (50.0%)
Technology and IT	29 (8.1%)	25 (18.5%)	4 (1.8%)	–
Healthcare	54 (15.2%)	17 (12.6%)	37 (16.9%)	–
Construction and real estate	13 (3.7%)	8 (5.9%)	5 (2.3%)	–
Trade and services	3 (0.8%)	2 (1.5%)	1 (0.5%)	–
Media	4 (1.1%)	3 (2.2%)	1 (0.5%)	–
Business and administration	23 (6.5%)	14 (10.4%)	9 (4.1%)	–
Transport and logistics	2 (0.6%)	1 (0.7%)	1 (0.5%)	–
Natural sciences	32 (9.0%)	16 (11.9%)	16 (7.3%)	–
Other	64 (18.0%)	23 (17.0%)	40 (18.3%)	1 (50.0%)

Note. Percentages may not sum to 100% due to rounding. [a]German school qualification levels. Dashes indicate no cases in that subgroup. Two participants did not report their gender.

Option B), (3) receiving brief moral advice, and (4) indicating how likely they would be to make the same decision again, as well as how helpful they rate the respective advice.

For each dilemma, participants received advice that was randomly framed as either human- or AI-generated and either supported or opposed their initial decision. Advisor and advice direction were randomized independently at the trial level. After completing all dilemmas, participants responded to the individual-difference questionnaires (AI aversion, wisdom, moral identity) followed by demographic questions. The study concluded with a debriefing that explained the scientific purpose of the experiment and clarified that all advice texts were standardized for research purposes, regardless of advisor framing. No time limits were imposed at any point during the procedure.

6.3.3 Measures

Moral Dilemmas

Participants completed four moral dilemmas: two everyday dilemmas adapted from Singer (2019) and two high-conflict sacrificial dilemmas adapted from Greene and colleagues (2004; 2008). A typical high-conflict example was the trolley dilemma: "A runaway trolley is heading toward five people. You can pull a lever to divert it onto another track, where it will kill one person instead. Would you pull the lever?" All dilemmas were translated into German, pilot-tested for clarity, and selected to avoid emotionally polarizing or politically charged content.

Advice Stimulus

After each initial decision, participants received a brief two-sentence piece of advice recommending one of the two options. The advice texts were identical across advisor-type conditions to isolate source effects. Advisor (AI vs. human) and advice direction (favoring Option A vs. Option B) were randomized independently for each dilemma, such that congruence or incongruence was defined relative to participants' initial choices. This yielded a fully within-subjects 2 (advisor type: AI vs. human) $\times$ 2 (advice direction: congruent vs. incongruent) factorial design.

In the AI condition, the advice was labelled as ChatGPT OpenAI (2024) output. In

the human condition, the advice was presented as a quotation attributed to a previous participant and was accompanied by a portrait image. To avoid privacy concerns, all faces were AI-generated portraits empirically shown to be indistinguishable from real photographs (Nightingale & Farid, 2022). Images were drawn from an openly accessible OSF dataset (https://osf.io/ru36d/) and were randomly assigned on each human-advice trial.

Dependent Variables

Before receiving advice, participants made a binary choice between Option A and Option B. After receiving the advice, they indicated how likely they were to repeat the same decision using a continuous slider from 1 (definitely Option A) to 101 (definitely Option B). This post-advice rating represented the continuous dependent measure, decision after advice (DAA).

Two binary variables were derived from the continuous response:

Decision consistency was coded as 1 if the post-advice decision matched the initial decision and 0 if it differed. Values from 1–50 were classified as Option A, values from 52–101 as Option B; the midpoint (51) was excluded because it represented an indeterminate position.

Advice following was coded only for incongruent trials. A value of 1 indicated that the participant shifted their decision toward the advice, and 0 indicated that they maintained their initial choice.

Individual-Difference Measures

Participants subsequently completed three validated questionnaires: (a) the AI Aversion Scale (Sindermann et al., 2021), measuring general attitudes toward AI; (b) the Moral Identity Scale (Aquino & Reed, 2002), assessing the centrality of moral traits to self-concept; and (c) the San Diego Wisdom Scale (SD–WISE; (Thomas et al., 2022)), measuring wisdom-related traits such as reflection and prosocial orientation. All scales were presented and german.

6.3.4 Data Analysis

Statistical Approach

All statistical analyses were conducted in R (Version 4.5.1, 2025-06-13) using RStudio (Version 2025.05.1+513). Data were imported from Excel files and managed using the `readxl` and

`openxlsx` packages, and pre-processed with functions from the `tidyverse` ecosystem, including `dplyr` and `tidyr`, for data cleaning, recoding, and the construction of derived variables. Generalized linear mixed-effects models and linear mixed-effects models were estimated with the `lme4` package, and model-based visualizations were created using `ggplot2`. Estimated marginal means and predicted probabilities for the graphical presentation and interpretation of interaction effects were obtained using the `ggemmeans` package. Missing responses were retained, as generalized linear mixed-effects models (GLMMs) naturally accommodate unbalanced repeated-measures data without listwise deletion (Hilbert et al., 2019). Statistical significance was evaluated at $\alpha = .05$ (two-tailed).

Model Specification

Binary outcomes (decision consistency, advice following) were analyzed using GLMMs with a logit link; continuous ratings (DAA) were analyzed using LMMs with Gaussian errors. All models included random intercepts for participants to account for repeated measurements. The general model structure was:

$$g(Y_{ij}) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{k,ij} + u_{0i} + \epsilon_{ij}, \tag{6.1}$$

where:

- $g(\cdot)$: link function (logit for GLMMs; identity for LMMs)
- Y_{ij}: outcome for participant i on dilemma j
- β_0: fixed intercept
- β_k: fixed effects for the K predictors
- $X_{k,ij}$: predictor values for participant i on dilemma j
- $u_{0i} \sim \mathcal{N}(0, \sigma_u^2)$: participant-level random intercept
- ϵ_{ij}: residual error term (included only in LMMs)

Random slopes were not included due to the small number of trials per participant (four dilemmas), resulting in insufficient data to estimate slope variance reliably; preliminary

random-slope models exhibited convergence issues. Continuous predictors (e.g., wisdom, moral identity) were grand-mean centered.

Model Diagnostics and Convergence

Models were fitted with the `bobyqa` optimizer using `maxfun = 2e5`. Convergence was assessed via gradient diagnostics. The initial model that attempted to predict the dichotomized post-advice decision exhibited quasi-complete separation, yielding extreme parameter estimates and unstable fits. Redefining the dependent variable as decision consistency resolved this issue and aligned the analysis with the theoretical goal of examining moral stability after advice (creating a new dependent binary variable that consists of the information if the decision after advice (DAA) matches the initial decision (when receiving opposing advice). We then used the advice as a predictor to test for differences between both advisors and advice types).

Residual inspection showed no evidence of overdispersion or systematic misfit in GLMMs, and Q–Q plots supported the normality assumption for LMM residuals.

Inference and Effect Size Reporting

Fixed effects in GLMMs were evaluated using Wald z tests, and coefficients were exponentiated to yield odds ratios (OR) with 95% confidence intervals. LMM results are reported as unstandardized coefficients (b) with corresponding t values. All estimated marginal means and contrasts were computed with Tukey-adjusted comparisons unless stated otherwise.

Post-hoc Sensitivity Analysis

Post-hoc power analyses using G*Power (Faul et al., 2007) indicated that both the everyday-dilemma effect (OR = 0.22, $N = 557$) and the sacrificial-dilemma effect (OR = 8.60, $N = 529$) were associated with a computed power of 1.00 at $\alpha = .05$. Given the near-zero baseline switching rates in sacrificial dilemmas, the large observed odds ratio reflects quasi-separation rather than a substantively large behavioral effect. Thus, the study possessed more than adequate sensitivity to detect effects of the magnitude estimated in the logistic mixed-effects models.

Data Availability & Ethical Approval Code

Ethical approval was obtained from the Ethics Committee of the University of Regensburg (approval code: 23-3326-101). The preregistration is available at https://osf.io/qd95s.

6.4 Results

6.4.1 Summary

Across analyses, participants generally maintained their initial moral decisions, but decision stability declined significantly when advice opposed their prior choice. In everyday dilemmas, opposing advice significantly increased switching behavior, whereas altruistic decisions showed only a marginal advantage in consistency. In sacrificial dilemmas, opposing advice markedly reduced stability for utilitarian decisions, whereas participants who initially decided deontological were largely unaffected by opposing advice. Across all contexts, both AI and human advisors exerted comparable influence, and advice perceived as more helpful was more likely to be adopted. In addition, continuous outcome analyses showed that even when categorical reversals were rare, particularly in sacrificial dilemmas, participants systematically shifted their post-advice evaluations in the direction of the advice, indicating that moral decisions remained sensitive to advisory influence at a continuous level despite high binary stability.

6.4.2 Mean Decision Patterns and Advice-Following Rates

Figure 6.1 illustrates the distribution of initial moral decisions across the four dilemmas. In the Footbridge dilemma, $n = 103$ participants initially selected the deontological option, whereas $n = 31$ chose the utilitarian option. In the Sacrifice dilemma, $n = 91$ participants chose the deontological option and $n = 52$ the utilitarian option. In the everyday dilemmas, $n = 97$ participants selected the altruistic option and $n = 48$ the egoistic option in the Neighbor dilemma, while $n = 90$ chose the altruistic option and $n = 57$ the egoistic option in the Homeless Man dilemma.

Advice-following rates in opposing trials are shown in Figure 6.2. In the Footbridge dilemma, $n = 10$ participants followed deontological advice, compared to $n = 5$ who followed utilitarian advice. In the Sacrifice dilemma, $n = 7$ participants followed deontological advice and $n =$

12 followed utilitarian advice. In everyday dilemmas, altruistic advice was followed more frequently than egoistic advice, with $n = 17$ versus $n = 11$ participants in the Neighbor dilemma and $n = 14$ versus $n = 9$ participants in the Homeless Man dilemma.

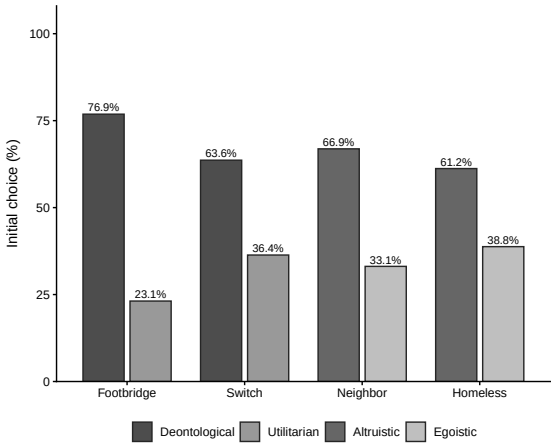


Figure 6.1: Initial moral decisions by dilemma. Bars show the percentage of participants choosing each option (deontological vs. utilitarian in sacrificial dilemmas; altruistic vs. egoistic in everyday dilemmas).

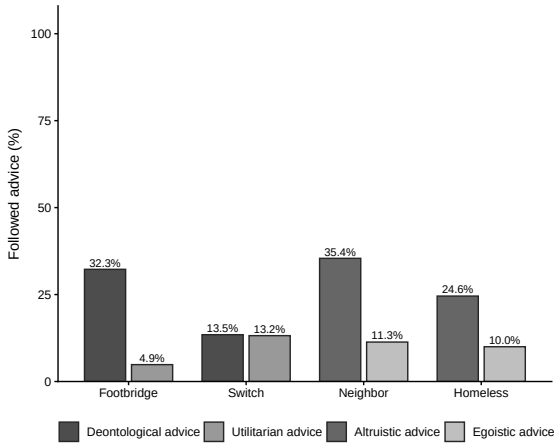


Figure 6.2: Advice-following rates in opposing trials by dilemma and advice direction. Bars indicate the percentage of participants who changed their initial decision to match the received advice when the advice opposed their prior decision.

6.4.3 Main Analysis: Decision Consistency Across Dilemma Types

To test how advice congruence and initial moral decision influenced the stability of moral decisions after receiving advice, two binomial generalized linear mixed-effects models (GLMMs) were estimated separately for everyday and sacrificial dilemmas. The dependent variable was decision consistency (1 = maintained initial decision, 0 = switched after advice). Each model included fixed effects for advice congruence (supporting vs. opposing), initial decision type (altruistic vs. egoistic in everyday dilemmas; deontological vs. utilitarian in sacrificial dilemmas), and their interaction, with random intercepts for participants. The full model estimates are presented in Table 6.2, and corresponding estimated marginal probabilities are displayed in Figure 6.3.

Table 6.2: Predictors of Decision Consistency in Everyday and Sacrificial Dilemmas

Predictor	**Estimate**	**SE**	z	p
Everyday dilemmas				
Intercept	2.65	0.58	4.56	**<.001**
Advice congruence (no vs. yes)	-1.51	0.53	-2.87	**.004**
Initial decision (altruistic vs. egoistic)	0.94	0.51	1.84	.067
Advice × Decision	0.80	0.70	1.15	.249
Sacrificial dilemmas				
Intercept	-11.32	2.02	-5.60	**<.001**
Advice congruence (no vs. yes)	2.15	1.13	1.91	.056
Initial decision (utilitarian vs. deontological)	-1.39	1.61	-0.86	.388
Advice × Decision	4.71	1.99	2.36	**.018**

Note. Generalized linear mixed-effects models with binomial logit link. Dependent variable: decision consistency (1 = consistent, 0 = switched). Bold p values indicate statistical significance at $p < .05$.

Everyday Dilemmas

As shown in Table 6.2, the model for everyday dilemmas revealed a significant main effect of advice congruence. Participants were substantially less likely to remain consistent with their initial decision when the received advice opposed their prior decision ($\beta = -1.51$, $SE = 0.53$, $z = -2.87$, $p = .004$, OR = 0.22, 95% CI [0.08, 0.63]).

The main effect of decision type indicated that altruistic decisions tended to be more stable than egoistic ones ($\beta = 0.94$, $SE = 0.51$, $z = 1.84$, $p = .067$, OR = 2.56, 95% CI [0.92, 7.10]), although this effect did not reach significance. The interaction between advice congruence and decision type was not significant ($\beta = 0.80$, $SE = 0.70$, $z = 1.15$, $p = .249$), suggesting that opposing advice reduced consistency to a similar degree for altruistic and egoistic decisions.

Figure 6.3a illustrates the estimated marginal probabilities of maintaining one's initial decision in everyday dilemmas as a function of advice congruence and decision type. When

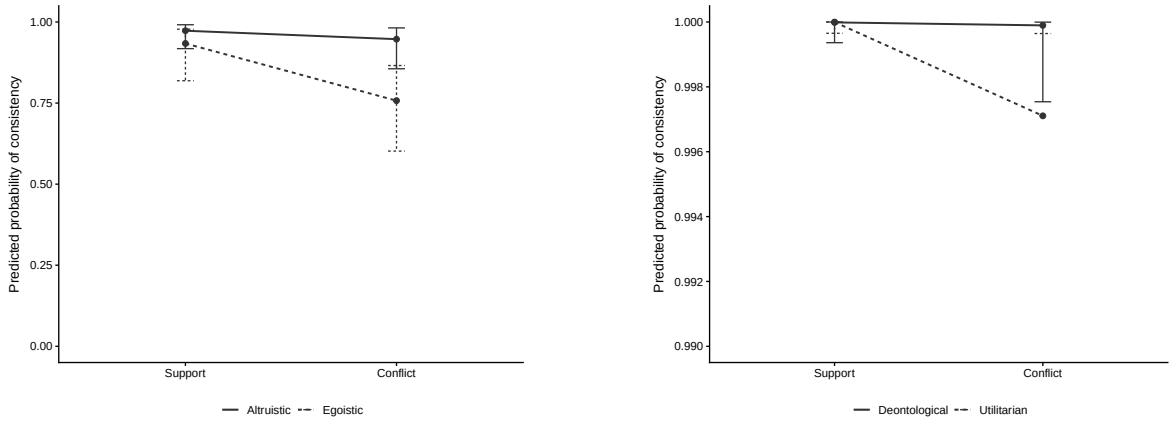
advice supported the initial decision, consistency was high for both egoistic ($p = .93$, $SE = .04$) and altruistic decisions ($p = .97$, $SE = .02$). Under opposing advice, consistency declined, particularly for egoistic decisions ($p = .76$, $SE = .07$), whereas altruistic decisions remained comparatively stable ($p = .95$, $SE = .03$). Relative to supportive egoistic advice, the odds of maintaining an egoistic decision under opposing advice were significantly reduced (OR = 0.22, $p = .012$), whereas no reliable reduction was observed for altruistic decisions (OR = 1.27, $p = .891$).

Sacrificial Dilemmas

As summarized in Table 6.2, the model for sacrificial dilemmas revealed a very low baseline probability of switching when advice supported deontological decisions ($\beta = -11.32$, $SE = 2.02$, $z = -5.60$, $p < .001$). A marginal main effect of opposing (vs. supporting) advice ($\beta = 2.15$, $SE = 1.13$, $z = 1.91$, $p = .056$) suggested a higher likelihood of switching when advice contradicted the initial decision. Crucially, the interaction between advice congruence and decision type was significant ($\beta = 4.71$, $SE = 1.99$, $z = 2.36$, $p = .018$), indicating that utilitarian decisions were especially unstable when faced with opposing advice.

Figure 6.3b shows the corresponding model-based predictions for sacrificial dilemmas. Across all conditions, consistency was near ceiling, indicating that participants rarely changed their initial moral decisions. When advice supported deontological decisions, the predicted probability of maintaining the initial decision was extremely high ($p = 1.2 \times 10^{-5}$), and remained similarly high under opposing advice ($p = 1.0 \times 10^{-4}$). For utilitarian decisions, consistency was likewise near ceiling under supportive advice ($p = 3.0 \times 10^{-6}$), but decreased slightly when advice opposed the initial decision ($p = 2.9 \times 10^{-3}$). This change corresponded to a large odds ratio relative to the supportive deontological baseline (OR = 238.85, $p = .001$), reflecting the very low base rate of switching in sacrificial dilemmas rather than a substantial absolute increase in reversals.

Because consistency rates in sacrificial dilemmas were near ceiling, the associated predicted probabilities cluster tightly around 1, which limits visual variation in the plot and reflects a floor effect in switching behavior rather than a lack of substantive differences between conditions.



(a) Everyday dilemmas.

(b) Sacrificial dilemmas.

Figure 6.3: Estimated marginal probabilities of decision consistency by advice congruence and initial decision type. Error bars represent 95% confidence intervals.

Across both dilemma types, participants overwhelmingly maintained their initial moral decisions when advice was supportive. However, opposing advice significantly reduced stability in everyday dilemmas and, especially, when utilitarian decision-making was challenged in sacrificial dilemmas. These results indicate that moral decisions are generally robust to congruent feedback but can shift under dissonant advice, consistent with the notion that advice can destabilize intuitive decisions when it contradicts one's prior moral stance.

Continuous Outcome Analyses to Elaborate Initial Findings

As shown in Table 6.2, sacrificial dilemmas exhibited a strong floor effect in the switching analysis, with participants rarely revising an initially deontological decision. To capture evaluative adjustments that did not cross the categorical decision threshold, linear mixed-effects models (LMMs) were estimated for the continuous post-advice ratings (see Table 6.3).

Across both dilemma types, advice exerted a robust influence on post-advice decisions. In everyday dilemmas, egoistic (vs. altruistic) advice substantially reduced post-advice ratings ($\beta = -45.10$, $SE = 4.64$, $t = -9.73$, $p < .001$), and this effect did not differ between human and AI advisors. In sacrificial dilemmas, utilitarian (vs. deontological) advice produced markedly higher post-advice ratings ($\beta = 34.45$, $SE = 4.20$, $t = 8.20$, $p < .001$), with no reliable interaction between advice and advisor type. These continuous decision shifts align with the patterns observed in the GLMM. Although categorical reversals were rare,

Table 6.3: Linear Mixed-Effects Model Predicting Continuous Post-Advice Decisions

Predictor	**Estimate**	**SE**	t	p
Everyday dilemmas				
Intercept	64.83	3.77	17.21	**<.001**
Advice (ego vs. altru)	−45.10	4.64	−9.73	**<.001**
Advisor (AI vs. human)	1.28	5.32	0.24	.810
Advice × Advisor	−1.12	6.69	−0.17	.867
Sacrificial dilemmas				
Intercept	43.38	3.55	12.22	**<.001**
Advice (util vs. deont)	34.45	4.20	8.20	**<.001**
Advisor (AI vs. human)	−8.71	5.45	−1.60	.111
Advice × Advisor	10.98	6.40	1.72	.087

Note. Dependent variable: post-advice decision-making (0–100 scale). Models include random intercepts for participants. Bold p values indicate statistical significance at $p < .05$. Advice coding: egoistic vs. altruistic (everyday); utilitarian vs. deontological (sacrificial). Advisor coding: AI vs. human.

particularly in sacrificial dilemmas, participants nonetheless showed systematic movement on the response scale in the direction of the received advice. The GLMM and LMM results indicate that sacrificial decisions are relatively stable in binary terms yet remain responsive to advice at a continuous evaluative level.

Advisor Influence on the Decision

Two linear mixed-effects models tested whether advice direction (deontological vs. utilitarian in sacrificial dilemmas; altruistic vs. egoistic in everyday dilemmas) and advisor type (human vs. AI) influenced post-advice moral decisions (DAA). In both contexts, advice direction exerted a strong main effect. In sacrificial dilemmas, utilitarian advice led to substantially lower DAA scores, indicating more utilitarian responding than deontological advice for both human, $\Delta = -34.5$, $SE = 4.24$, $t(260) = -8.12$, $p < .001$, and AI advisors, $\Delta = -45.4$, $SE = 4.87$, $t(265) = -9.33$, $p < .001$. Advisor type did not significantly affect DAA scores within advice

conditions ($p = .499$, $p = .114$). Similarly, in everyday dilemmas, egoistic advice produced markedly lower DAA values (reflecting more egoistic responding) than altruistic advice, for both human, $\Delta = -45.1$, $SE = 4.67$, $t(280) = -9.66$, $p < .001$, and AI advisors, $\Delta = -46.2$, $SE = 4.78$, $t(284) = -9.67$, $p < .001$. Again, advisor type showed no reliable differences ($p = .968$, $p = .812$). Overall, participants consistently adjusted their moral decisions in the direction of the advice received, with comparable magnitude for AI and human advisors.

Table 6.4: Estimated Marginal Means (DAA) by Advice Direction and Type in Opposing-Only Trials

		Sacrificial Dilemmas			**Everyday Dilemmas**		
Advisor	**Advice**	*M*	*SE*	95% CI	*M*	*SE*	95% CI
Human	Deontological / Altruistic	43.4	3.58	[36.3, 50.4]	64.8	3.79	[57.4, 72.3]
Human	Utilitarian / Egoistic	77.8	2.37	[73.2, 82.5]	19.7	2.83	[14.2, 25.3]
AI	Deontological / Altruistic	34.7	4.22	[26.4, 43.0]	66.1	3.83	[58.6, 73.6]
AI	Utilitarian / Egoistic	80.1	2.45	[75.3, 84.9]	19.9	2.92	[14.1, 25.6]
Simple effects of advice within advisor type							
Human	Δ (Advice)	-34.5, $t(260) = -8.12$, $p < .001$			-45.1, $t(280) = -9.66$, $p < .001$		
AI	Δ (Advice)	-45.4, $t(265) = -9.33$, $p < .001$			-46.2, $t(284) = -9.67$, $p < .001$		

Note. Higher DAA values indicate more deontological responding in sacrificial dilemmas and more egoistic responding in everyday dilemmas. Models include random intercepts for participants. Advice effects reflect contrasts between deontological/altruistic versus utilitarian/egoistic advice within each advisor type.

Advice-following Behavior. To quantify behavioral susceptibility to external moral advice, we examined the proportion of participants who changed their initial decision to match the advice when the two opposed. Table 6.5 summarizes follow rates across advisor types, advice directions, and dilemmas. Overall, participants were considerably more likely to follow altruistic or deontological advice than egoistic or utilitarian advice. On average, approximately one quarter of participants accepted prosocial or deontological advice, compared to roughly 10% who accepted self-serving or utilitarian advice. Follow rates were slightly higher for human advisors (particularly for altruistic advice; 34.6%) than for AI advisors (24.5%), although this

difference was not pronounced and did not reach significance.

Across dilemmas, advice-following was most frequent in everyday contexts (up to 35.4% for altruistic advice) and least frequent for utilitarian advice in high-conflict sacrificial dilemmas (4.9%). These results suggest that participants were selectively receptive to advice promoting moral norms of care and deontological consistency, while advice encouraging utilitarian trade-offs or egoistic benefits was rarely adopted.

Predictors of Advice Following. To examine individual and situational factors predicting advice compliance, a generalized linear mixed-effects model (GLMM) with a logit link was fitted on the binary outcome advice-following (yes/no). The model included centered trait-level predictors Wisdom and Moral Identity–Internalization, the between-subject factor advisor (human vs. AI), and the trial-level predictor perceived helpfulness, with random intercepts for participants. Results indicated that only perceived helpfulness significantly predicted the likelihood of following advice, $b = 0.04$, $SE = 0.01$, $z = 6.27$, $p < .001$. Participants were more likely to follow advice they perceived as helpful. In contrast, neither wisdom ($p = .68$), internalization of moral identity ($p = .14$), nor the advisor type ($p = .50$) had a significant influence on advice following. The random intercept variance ($\sigma^2 = 0.26$) suggested moderate between-person variability in baseline compliance. Overall, these results indicate that advice-following was driven primarily by situational perceptions of helpfulness rather than by individual differences or the advisor.

Perceived Helpfulness of Advice. To examine whether perceived helpfulness varied as a function of advisor type and advice direction, two linear mixed-effects models were fitted for the high-conflict and everyday dilemmas separately, each including fixed effects for advisor (human vs. AI), advice type, and their interaction, with random intercepts for participants and dilemmas. For high-conflict dilemmas, the model revealed a significant advisor × advice type interaction, $b = -14.09$, $SE = 6.18$, $t(216) = -2.28$, $p = .024$. Simple effects showed that advice from human advisors was perceived as more helpful overall, $b = 11.44$, $SE = 5.25$, $p = .031$, but this advantage was specific to deontological advice. When providing utilitarian advice, human advisors were rated as less helpful than AI advisors (see Figure 6.4). For everyday dilemmas, the analysis indicated a significant main effect of advice direction,

Table 6.5: Advice-Following Rates by Advisor Type and Dilemma Type

(A) Advisor Type and Advice Direction

Advisor	**Advice Direction**	n	n_{follow}	*% Followed*
Human	*Total*	**294**	**49**	**16.7**
	Altruistic	52	18	34.6
	Deontological	47	11	23.4
	Egoistic	95	10	10.5
	Utilitarian	100	10	10.0
AI	*Total*	**275**	**36**	**13.1**
	Altruistic	53	13	24.5
	Deontological	36	6	16.7
	Egoistic	92	10	10.9
	Utilitarian	94	7	7.5

(B) Dilemma Type and Advice Direction

Dilemma	**Advice Direction**	n	n_{follow}	*% Followed*
Sacrificial – Footbridge	Deontological	31	10	32.3
	Utilitarian	103	5	4.9
Sacrificial – Switch	Deontological	52	7	13.5
	Utilitarian	91	12	13.2
Everyday – Helping	Altruistic	48	17	35.4
	Egoistic	97	11	11.3
Everyday – Donation	Altruistic	57	14	24.6
	Egoistic	90	9	10.0

Note. Totals in panel (A) aggregate across advice directions within each advisor type. Values reflect mismatching (opposing) trials only. Higher percentages indicate stronger tendencies to change one's initial decision in the direction of the received advice.

$b = -20.85$, $SE = 4.36$, $t(265) = -4.78$, $p < .001$, with egoistic advice perceived as less helpful than altruistic advice. Neither the main effect of advisor ($p = .85$) nor the interaction ($p = .84$) reached significance. Thus, while moral advice content strongly influenced perceived helpfulness in both contexts, only in sacrificial dilemmas did the advisor's identity moderate this effect.

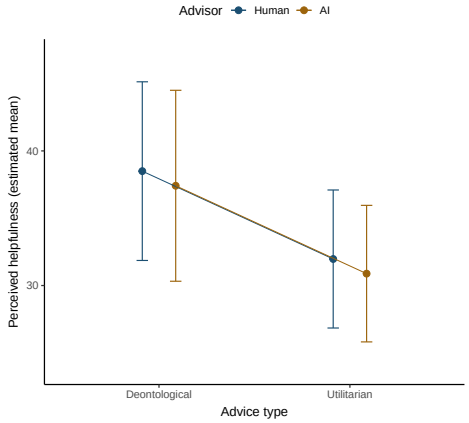


Figure 6.4: Estimated marginal means of perceived helpfulness by advisor type and advice direction in high-conflict dilemmas. Error bars represent 95% confidence intervals.

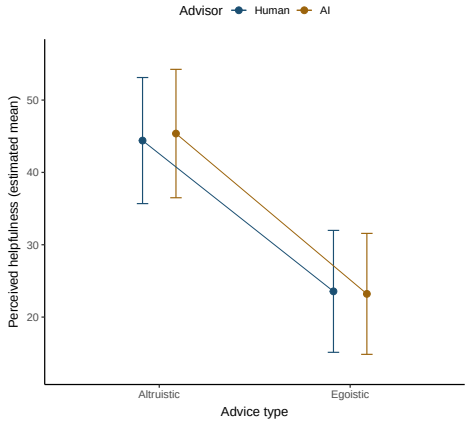


Figure 6.5: Estimated marginal means of perceived helpfulness by advisor type and advice direction in everyday dilemmas. Error bars represent 95% confidence intervals.

6.5 Discussion

6.5.1 Summary of Key Findings

The present study investigated whether moral advice can retrospectively alter prior moral decisions, and whether such influence differs between human and artificial advisors. We examined four dilemmas which consisted of two everyday moral and two sacrificial dilemmas. Across all dilemmas, participants generally maintained their initial moral stance, yet decision stability declined significantly when the received advice opposed the initial decision. This pattern confirms our preregistered hypothesis that post-decisional moral advice can shape subsequent moral evaluations even after individuals have committed to a decision. Importantly, the persuasive effect was driven primarily by the congruence and normative content of the advice rather than by its advisor, as human and AI advisors exerted comparable influence. Moreover, continuous outcome analyses showed that participants' evaluative decisions shifted

reliably in the direction of the advice, indicating that moral decisions can remain covertly malleable even when categorical reversals are rare.

6.5.2 Post-Decisional Advice Shifts Moral Decisions Asymmetrically

Everyday Dilemmas

In everyday moral conflicts, participants were significantly more likely to revise their initial decision when the advice opposed their first choice. This instability was slightly asymmetric: those who initially chose the egoistic option, thus deviating from the socially preferred altruistic norm, were more responsive to opposed advice promoting altruism, although this asymmetry did not reach statistical significance. In such cases, the advice appears to function as a social corrective, facilitating realignment with communal standards of generosity and care. In contrast, when participants had initially selected the socially endorsed altruistic option, advice encouraging egoistic behavior had little impact. Such "less social" advice was easily devalued or dismissed, reducing cognitive dissonance and preserving moral self-consistency. Nevertheless, both types of contradictory advice resulted in a significant reduction in decision consistency, underscoring the destabilizing effect of dissonant advice.

Several psychological mechanisms likely operate simultaneously in these scenarios. When participants receive congruent advice, it is unsurprising that their initial and post-advice decisions align. This stability reflects well-documented confirmation tendencies: individuals selectively seek, value, and integrate information consistent with their prior beliefs or choices (Bailey et al., 2023; Ditto et al., 2009; Schultze et al., 2017). Advice that supports one's initial decision reinforces coherence and reduces uncertainty, thereby sustaining the perception of decisional validity and moral integrity.

When receiving contradicting advice, however, participants are likely to experience cognitive tension, as the new information undermines the initial moral decision. This conflict between an existing belief and incongruent advice constitutes a form of cognitive dissonance (Festinger, 1962). Such dissonance can be resolved through two opposing strategies: either by revising one's decision to accommodate the new input or by devaluing the opposing information to preserve internal consistency (Ditto et al., 2009; Festinger, 1962). The direction of adjustment depends critically on the social desirability and normative weight of the alternative moral

position. When participants initially decide egoistically and subsequently receive altruistic advice, the advice operates as a socially sanctioned corrective. Even though moral dilemmas lack objective right or wrong answers, altruism typically represents a culturally endorsed and morally prototypical stance (Aquino & Reed, 2002; Janoff-Bulman et al., 2009). Hence, the initial egoistic decision elicits dissonance both because it conflicts with internalized moral norms and because the corrective advice highlights this incongruity. The combined moral and social tension increases the motivational pressure to realign one's stance toward the altruistic, socially accepted alternative. Conversely, when participants initially choose the altruistic option and are confronted with egoistic advice, the contradiction appears less threatening. Because altruistic reasoning aligns with socially desirable moral schemas and self-definitional moral identity (Aquino & Reed, 2002; Monin & Jordan, 2009), the devaluation of norm-deviant, egoistic advice is relatively effortless. In this case, dismissing the advice not only reduces cognitive dissonance but also preserves a sense of moral integrity and social desirability. As Haidt (2001) and Haidt and Joseph (2004) argued, moral reasoning often functions post-hoc to defend intuitively grounded moral positions, while Tetlock (2002) and Cialdini and Goldstein (2004) showed that impression-management concerns further stabilize conformity to socially approved norms. These mechanisms explain why opposing advice consistently reduces decision stability yet does so asymmetrically: socially corrective, altruistic cues are more persuasive, whereas socially deviant, egoistic cues are more easily dismissed. Although categorical changes were limited, participants' continuous ratings shifted in line with the advice, suggesting underlying evaluative movement even when explicit choices remained stable.

Sacrificial Dilemmas

A similar but more polarized asymmetry emerged in the high-conflict sacrificial dilemmas. Participants who initially made utilitarian decisions (often endorsing harm to achieve the greater good) were substantially more likely to revise their decision when confronted with deontological advice, whereas those who had chosen deontologically displayed near-complete resistance to utilitarian counter-advice. This divergence underscores that utilitarian reasoning, although cognitively deliberate, is more susceptible to socially corrective influence, while deontological decisions remain affectively entrenched and normatively self-reinforcing. Deontological advice

thus served as an emotionally grounded anchor that reaffirmed the deeply internalized moral prohibition against harm and offered participants a socially legitimate rationale to reverse their decision. In contrast, advice advocating utilitarian trade-offs conflicted with prevailing harm-avoidance norms and was easily discounted.

These results extend the dynamics observed in everyday moral contexts to dilemmas involving physical harm, suggesting that moral correction operates more forcefully when the normative violation is emotionally salient. When advice was congruent with participants' initial decision, decisions remained stable. This is consistent with confirmation processes that favor coherence and self-consistency (Bailey et al., 2023; Ditto et al., 2009; Schultze et al., 2017). Yet, when advice contradicted participants' decision, the moral dissonance became pronounced. As in everyday contexts, participants appeared to manage this tension by either re-evaluating their stance or by devaluing the advice. However, in sacrificial contexts, these mechanisms were amplified by the emotional intensity of the moral conflict. Cognitive dissonance (Festinger, 1962) was not merely epistemic but moral in nature, threatening one's identity as a non-harmful agent and activating self-protective processes identified in motivated reasoning research (Ditto et al., 2009).

From a dual-process perspective, this pattern aligns with evidence that deontological decisions are grounded in automatic affective responses, whereas utilitarian reasoning relies on slower, controlled cognition (J. D. Greene et al., 2004, 2008). Controlled processes are more flexible and thus more vulnerable to norm-congruent correction when opposing emotional or social cues are introduced. By contrast, deontological decisions, anchored in aversive emotional reactions to harm, are more resistant to revision. Coherence-based reasoning theories (Holyoak, 1999; D. Simon et al., 2004) further suggest that belief networks reorganize dynamically to minimize internal conflict; when the incoming information aligns with dominant cultural norms (such as prohibitions against instrumental harm) the cognitive system reorganizes toward deontological coherence.

Deontological advice therefore functions as a socially corrective cue, reinforcing the cultural expectation to avoid direct harm and restoring alignment with internalized moral values. Consistent with Janoff-Bulman et al. (2009), prescriptive cues emphasizing moral restraint evoke less defensiveness and more willingness to reconsider than norm-violating or harm-endorsing

messages. The corrective impact of such advice also reflects impression-management and social conformity motives (Cialdini & Goldstein, 2004; Tetlock, 2002): by endorsing deontological restraint, individuals can maintain a positive moral image and reaffirm commitment to shared moral norms (Aquino & Reed, 2002). In this sense, deontological advice plays a role analogous to altruistic advice in everyday dilemmas, faciliating not only cognitive consonance but also social and affective coherence. As proposed by Haidt (2001) and Haidt and Joseph (2004), moral reasoning thus serves less to uncover objective truths than to preserve social harmony and moral legitimacy. It is important to note, however, that despite the near absence of categorical reversals in sacrificial dilemmas, participants' continuous evaluations nonetheless shifted reliably in the direction of the advice, indicating that even ostensibly stable deontological decisions remain susceptible to more subtle forms of moral influence.

6.5.3 Advisor: Human and AI Advice Exert Comparable Influence

Across both dilemma contexts, the advisor's identity did not significantly affect the magnitude or direction of post-decisional change. Participants adjusted their decisions equally in response to AI and human advice, supporting the preregistered hypothesis that AI advice can be as persuasive as human advice. This equivalence suggests that once advice is framed as competent and contextually relevant, individuals rely more on its informational value than on its advisor per se. The result extends prior evidence showing that AI advice influences moral decisions even when participants are explicitly aware of its artificial origin and tend to underestimate their own susceptibility (Krügel et al., 2022, 2023b, 2025; Leib et al., 2021, 2023).

From a theoretical standpoint, these findings align with the "fit model" of advice-taking proposed by Baines et al. (2024), which posits that persuasive strength depends on the alignment between advisor characteristics, task demands, and decision-maker expectations rather than on advisor type. When the informational content of advice fits the moral context (such as clear, binary advice in structured dilemmas) AI advisors appear to achieve comparable legitimacy to human advisors. This functional equivalence resonates with meta-analytic evidence indicating that perceived competence, helpfulness, and task relevance are the primary predictors of advice weighting (Bailey et al., 2023).

Importantly, the absence of differential effects also complements recent findings that AI persuasion operates robustly even under transparency or accountability manipulations. Studies by Leib et al. (2021, 2023) demonstrated that AI-generated advice promoting dishonesty increased rule-breaking behavior to the same extent as human-written messages, while honesty-promoting advice showed limited corrective impact. Similarly, Krügel et al. (2023b, 2025) found that ChatGPT's inconsistent moral recommendations shifted participants' moral decisions regardless of whether its nonhuman source was disclosed. These results imply that AI advice is processed as a socially valid input rather than an epistemically inferior one. The observed parity between human and AI advisors may further reflect the growing normalization of AI systems in decision-making domains (Dietvorst et al., 2015; Logg et al., 2019). As users increasingly interact with AI in everyday contexts, trust in algorithmic guidance becomes dynamic and context-dependent (Araujo et al., 2020; Ashoori & Weisz, 2019; Gaube et al., 2021; Kelly et al., 2023). Under such conditions, participants may not evaluate AI advice as categorically distinct from human recommendations but rather as another credible information source within a broader social-cognitive frame. Consequently, AI and human advisors converge in persuasive potential, provided the advice is perceived as competent, normatively grounded, and situationally appropriate.

Taken together, these findings contribute to the growing evidence that AI persuasion in moral contexts does not depend on anthropomorphic framing or explicit social trust but on the perceived diagnostic value of the information itself. When AI-generated advice fits the moral and social parameters of a decision, it can exert influence comparable to that of human advice, further challenging the assumption that moral legitimacy necessarily requires human agency.

6.5.4 Content Dominance: Prosocial and Deontological Advice as Universal Social Correctives

Across contexts, advice promoting altruistic or deontological positions exerted stronger influence than egoistic or utilitarian advice. This cross-domain asymmetry highlights that moral persuasion depends primarily on normative alignment rather than dilemma type. Prosocial and deontological advice provides a culturally legitimate rationale for revision,

thereby reducing dissonance and restoring moral coherence, whereas self-serving or harm-justifying cues threaten moral identity and are readily dismissed. This pattern is consistent with recent findings showing that even when moral advice is delivered by an artificial agent, rule-based deontological advice is relatively more effective than virtue- or role-based appeals, suggesting that normative clarity rather than advisor identity drives persuasive impact (B. Kim et al., 2021). These findings integrate the observed asymmetries across everyday and sacrificial dilemmas, underscoring that socially corrective advice consistently prevails over norm-deviant alternatives (Aquino & Reed, 2002; Ditto et al., 2009; Festinger, 1962; Janoff-Bulman et al., 2009).

6.5.5 Mechanism: Perceived Helpfulness as Proximal Driver of Moral Updating

Perceived helpfulness emerged as the only significant predictor of advice-following, whereas stable traits such as wisdom and moral identity internalization had no effect. This indicates that moral persuasion is driven less by enduring dispositions than by situational appraisals of relevance and utility. Participants adopted advice when it was perceived as helpful or applicable to the specific dilemma, independent of the advisor's nature and participants' general attitudes toward AI. This pattern suggests that moral updating depends on immediate judgments of informational fit rather than on stable trust dispositions or reflective virtues. These findings converge with recent evidence that perceived competence and usability govern AI reliance (Dressel & Farid, 2018; Ng et al., 2021). From an applied perspective, enhancing perceived helpfulness may represent the most direct and ethically sustainable lever for fostering reflective engagement with moral advice, whether human or AI.

6.5.6 Implications

The present findings carry several theoretical, practical, and ethical implications for the study and application of moral advice-taking.

Theoretical Implications.

The results extend existing theories of moral cognition and advice-following in three ways. First, they demonstrate that moral change can occur even after an initial decision, challenging

the assumption that moral decisions are stable once formed. Yet such change is slightly asymmetric and in tendencies norm-dependent: advice reinforcing socially sanctioned norms facilitates moral realignment, whereas norm-deviant cues are selectively resisted, when the advice does not match the initial decision. Second, the findings bridge dual-process and coherence-based accounts of moral reasoning. Deontological resilience reflects the affective entrenchment of intuitive harm-avoidance norms, while utilitarian fragility under opposing advice illustrates cognitive reorganization toward coherence when socially corrective input is introduced (J. D. Greene et al., 2004, 2008; Holyoak, 1999; D. Simon et al., 2004). Third, the observed parity between AI- and human advice refines current models of AI appreciation and aversion (Dietvorst et al., 2015; Logg et al., 2019) and supports current findings, that artificial and human advisors exert comparable influence when advising moral dilemmas. Therefore, moral persuasion appears to depend less on anthropomorphic trust than on the perceived legitimacy and normative alignment of the moral message itself.

Practical and Ethical Implications.

From an applied perspective, the finding that AI advice can influence moral decisions as effectively as human advice underscores the societal relevance of these mechanisms. Importantly, moral dilemmas in real life rarely unfold as abstract decisions; they often involve an initial decision that is later reconsidered in light of external input. By modeling post-decisional moral advice, the present design thus offers a step toward greater external validity and situational realism in the study of moral persuasion. This approach enhances practical relevance by mirroring how individuals encounter advice in everyday and institutional contexts after, rather than before, they have committed to a stance. Developers of moral advisory systems should therefore prioritize transparency, contextual fit, and perceived helpfulness to ensure that AI influence promotes informed reflection rather than passive conformity (Dressel & Farid, 2018; Ng et al., 2021). Given that prescriptive, prosocial advice is disproportionately persuasive, system designers and policymakers bear responsibility to align such advice with democratic values and moral pluralism. At the same time, the capacity of moral advice to correct norm-deviant reasoning highlights opportunities for constructive moral nudging in education, deliberative democracy, and public discourse, provided that autonomy, diversity

of moral perspectives, and reflective agency are preserved. Ultimately, the ethical challenge lies not in suppressing AI persuasion but in channeling it toward fostering moral awareness, critical engagement, and genuinely autonomous moral reasoning.

6.5.7 Limitations and Future Directions

While the present study advances understanding of moral advice-taking, several limitations warrant consideration and suggest directions for future research.

First, although the post-decisional design substantially enhances external validity by reflecting how people reconsider prior moral decisions in everyday contexts, the dilemmas remained hypothetical and consequence-free. This limits the emotional realism of moral conflict. Future studies could employ immersive or consequential paradigms, such as virtual reality, incentivized moral trade-offs, or ecologically grounded narratives to examine whether the observed asymmetries generalize to high-stakes or embodied contexts.

Second, the study deliberately controlled for advice content across human and AI advisors to isolate source perception effects. While this design choice strengthens internal validity, it also omits manipulation checks for perceived humanness, competence, or moral legitimacy. Consequently, it remains unclear whether participants consciously differentiated between AI and human advisors or relied primarily on message content. This omission represents a methodological limitation, as the absence of a manipulation check prevents firm conclusions about participants' actual advisor perception. Subsequent future research should therefore include explicit measures of advisor perception and perceived trustworthiness to test their mediating role in moral persuasion.

Third, the sample consisted primarily of university students from a Western cultural context. Although this homogeneity supports internal consistency and interpretability, it restricts generalizability across cultures and moral frameworks. Future research should therefore replicate the design across diverse cultural and demographic populations, where norms surrounding altruism, deontology, and trust in AI may differ substantially.

Future studies should further examine how stable individual differences moderate susceptibility to moral advice, potentially explaining why some individuals resist social correction while others readily adjust.

Finally, the present design involved single-trial exposures and unidirectional advice. Real-world moral communication, however, often unfolds through interactive and iterative exchanges. Future work should explore dynamic advisory settings in which AI or human systems adapt to users' moral reasoning over time, enabling analysis of reciprocal moral influence. Moreover, integrating personality- or trait-level moderators (such as trust in AI, moral conviction, or metacognitive insight) could further clarify interindividual variability in susceptibility to moral persuasion.

6.6 Conclusion

This study examined whether post-decisional moral advice from human and AI advisors can retrospectively influence moral decisions. By comparing everyday and sacrificial moral contexts, it explored how advice direction and advisor shape the stability and malleability of moral decisions after an initial decision. The present research demonstrates that moral decisions remain malleable even after initial decision and that external advice can retrospectively shape moral reasoning. Yet this influence is asymmetrical: advice that aligns with socially endorsed norms, such as altruism and harm avoidance, exerts a corrective force, whereas norm-deviant cues are resisted or devalued more often. These dynamics highlight the dual nature of moral advice as both a mechanism of coherence restoration and a socially corrective force, that promotes moral alignment yet risks fostering conformity at the expense of moral autonomy.

Across both dilemma types, the findings converge on the notion that moral advice functions primarily as a social corrective mechanism. Individuals strive to maintain coherence among their decisions, self-concept, and perceived moral integrity (Ditto et al., 2009; Festinger, 1962). When advice supports socially approved norms, it provides a legitimate avenue to restore consonance and moral self-consistency, whereas norm-deviant advice threatens moral self-regard and is devalued to protect one's identity (Monin & Jordan, 2009). In this sense, prescriptive advice that affirms prosocial or harm-avoidance norms offers a psychologically acceptable path to moral realignment (Aquino & Reed, 2002; Janoff-Bulman et al., 2009). From a broader social perspective, moral advice also engages impression-management motives: individuals anticipate social evaluation and align judgments with shared expectations (Cialdini & Goldstein, 2004; Tetlock, 2002). The persuasive power of moral advice therefore lies not in

its informational origin but in its alignment with the moral norms that define the social self. The findings raise critical questions about how moral reflection, authenticity, and responsibility are negotiated in technologically mediated contexts. Future research should aim to balance the benefits of reflective moral correction with the ethical imperative to preserve individual moral agency. Ultimately, the challenge lies not in resisting moral influence altogether but in cultivating systems and contexts that foster deliberate, autonomous, and socially responsive moral reasoning.

7 General Discussion

7.1 Summary of Main Findings

Across four empirical studies, this dissertation systematically examined how AI-generated moral advice, compared with advice from human advisors, influences moral decision-making across diverse contexts, dilemma structures, and advice timings. Despite substantial variation in design, a set of robust and convergent empirical patterns emerged.

Across all studies, the content and normative direction of moral advice reliably predicted moral decisions, whereas the advisor comparison revealed no consistent effect. In several contexts, deontological and altruistic advice exerted stronger influence than utilitarian or egoistic advice, regardless of advisor type or modality. This pattern held across classical sacrificial dilemmas, everyday interpersonal conflicts, conflicting advisory contexts, and post-decisional decision scenarios. Thus, persuasive impact depended overwhelmingly on the normative framing of the advice rather than on "who" or "what" delivered it.

Artificial and human advisors exerted comparable influence on moral decision-making in all four studies. Whether participants received a single piece of advice (Studies 1 and 2), jointly convergent human–AI advice (Study 3), directly conflicting advice from human and AI advisors (Study 3), or post-decisional advice (Study 4), acceptance rates did not differ systematically as a function of advisor type. Even in conditions designed to heighten interpersonal cues (e.g., photographic depictions of human advisors), AI did not suffer a persuasive disadvantage. These findings provide strong support for the irrelevance of the advisor in moral persuasion once advice is perceived as relevant and legitimate.

Across all studies, external moral advice produced meaningful shifts in moral decision-making. In pre-decisional settings, advice reliably increased the likelihood of selecting the recommended option (Studies 1–3). In post-decisional settings, advice reduced decision stability and prompted retrospective change (Study 4), demonstrating that even previously

formed moral decisions remain malleable in the presence of external advice.

Susceptibility to advice was notably asymmetrical and norm-dependent. Across both everyday and sacrificial dilemmas, normatively aligned advice (altruistic, deontological) exerted substantially stronger influence than norm-deviant advice (egoistic, utilitarian). Advice that supported socially undesirable or norm-violating options was often resisted, whereas advice aligned with socially endorsed standards frequently operated as a corrective, increasing the likelihood of shifts toward more prosocial or harm-avoidant choices, both before and after an initial decision.

Individual difference measures (e.g., moral identity, wisdom) yielded inconsistent effects across studies. By contrast, situational appraisals (particularly perceived helpfulness) emerged as stronger and more reliable predictors of advice-following, especially when participants were confronted with conflicting advice. Advice also modulated subjective experience: it increased perceived ease in high-conflict dilemmas and reduced perceived responsibility in lower-conflict contexts, although these effects were modest and less consistent than the robust behavioral shifts observed throughout the research program.

7.1.1 Brief Summary of Individual Studies

Study 1. Study 1 demonstrated that moral advice significantly influences decision-making across dilemmas with all levels of difficulty. Deontological advice reliably increased deontological decisions, whereas utilitarian advice showed comparatively limited persuasive power. Crucially, the advisor did not systematically alter its impact. Individual differences in wisdom predicted greater reliance on advice and higher perceived certainty, while advice itself increased perceived ease in difficult dilemmas and reduced perceived responsibility in less severe cases. These findings suggest that advice functions as a cognitive support mechanism under conditions of moral strain.

Study 2. Study 2 extended these results to everyday moral conflicts and showed that both artificial and human advice significantly shaped moral decisions. Altruistic advice was followed more frequently than egoistic advice, independent of advisor type. Social proximity influenced baseline decision tendencies but did not eliminate the effect of advice. Internalized moral values

were associated with greater advice-following, while subjective experience was shaped primarily by individual dispositions rather than by advisor, overall demonstrating the robustness of advice effects in ecologically relevant, relational contexts.

Study 3. Study 3 examined situations in which participants received either convergent or conflicting advice from a human and an artificial advisor simultaneously. When both advisors recommended the same option, advice adherence was particularly strong, indicating an additive persuasive effect. When advice conflicted, participants showed no systematic preference for either the human or the AI advisor. Notably, this study provides novel experimental evidence that human moral advice does not reliably override artificial advice even under conditions of direct conflict. Instead, advice-following was driven primarily by subjective appraisals such as perceived helpfulness and situational relevance. While similarity cues occasionally favored the human advisor, advisor identity remained secondary to content and evaluation. These findings challenge the assumption that human moral authority naturally supersedes AI advice and further support the principle of advisor irrelevance in moral persuasion.

Study 4. Study 4 investigated whether post-decisional moral advice can retrospectively alter already formed moral decisions. Across both everyday and sacrificial dilemmas, decision stability declined markedly when advice opposed the participant's initial decision. Artificial advice exerted an influence comparable in magnitude to that of human advice, prompting similar degrees of decision revision. Beyond demonstrating the malleability of post-decisional moral decision, these findings show that even initially confident, normatively aligned decisions can be revised when confronted with artificial advice. This underscores the capacity of AI not only to shape future moral decisions but also to retrospectively influence and reorganize existing moral decisions.

The four studies converge on a clear empirical pattern: moral advice is consistently persuasive; its influence is driven primarily by its normative content rather than by its advisor; artificial advisors are as effective as human advisors in shaping moral decisions; and moral decisions remain revisable even after initial decision. Across all contexts, normatively aligned advice exerts the strongest and most stable influence. These findings provide the empirical foundation for the subsequent theoretical interpretation, which conceptualizes AI-generated

moral advice as a cognitively efficient heuristic that reduces effort in morally demanding situations.

7.2 Empirical Synthesis and Cross-Study Integration

Across four empirical studies a consistent pattern emerged: moral advice, functions as an effort-reducing heuristic that reliably shapes moral decision-making. This section integrates the findings by identifying the overarching psychological mechanisms that account for the persuasive influence of both artificial and human advisors.

7.2.1 Advice as an Effort-Reducing Heuristic Across Contexts

The central unifying result of this research program is that moral advice serves as a cognitive shortcut. Independent of dilemma type, advisor identity, or temporal position, participants consistently relied on external advice as heuristic cues that simplified complex trade-offs, reduced deliberative effort, and provided a salient anchor for decision-making.

In Study 1, advice increased perceived ease in high-conflict dilemmas and reduced perceived responsibility in lower-conflict scenarios, indicating reduced cognitive and motivational burden. Study 2 demonstrated comparable effects in everyday interpersonal contexts. Study 3 showed that the heuristic function of advice demolishes, when opposing advice are presented simultaneously. What points toward the idea of heuristic use is, that human advice failed to overwrite artificial advice, indicating the irrelevance of the advisor and underscoring the heuristic use of cues. Study 4 demonstrated that advice can exert measurable influence even after an initial moral decision has been formed, indicating that heuristic uptake is sufficiently strong to override prior decisions. Notably, perceived helpfulness emerged as the only significant predictor of retrospective advice adoption, suggesting that participants integrated advice to the extent that it promised cognitive efficiency. In this sense, helpful advice may operate as an effort-reducing heuristic: it lowers cognitive load, facilitates coherence restoration, and provides a readily accessible rationale for revising one's prior decision. These findings support the broader interpretation that AI advice is often processed not as deep normative guidance but as a cognitively economical cue that assists in managing post-decisional conflict.

These findings support the interpretation that moral advice is predominantly processed via

the peripheral route of the Elaboration Likelihood Model (Petty & Cacioppo, 1986), particularly under conditions of limited motivation or cognitive capacity. Under such circumstances, individuals rely on simple, externally provided cues rather than engaging in systematic moral reasoning. Across all studies, advice functioned as precisely such as low-effort, high-efficiency cue. We further elaborate on this idea in Chapter 7.3.

7.2.2 Normative Asymmetry: Why Prosocial and Deontological Advice Prevails

A second consistent pattern across all studies was the disproportionate persuasive power of altruistic and deontological advice relative to egoistic and utilitarian advice. This normative asymmetry was evident in pre-decisional advice (Studies 1 and 2) and post-decisional contexts (Study 4). Participants more readily adopted advice that aligned with socially sanctioned moral norms (care, fairness, and harm avoidance)(Graham et al., 2013) while resisting advice that legitimized norm violations or self-interested outcomes.

This asymmetry can be understood through three converging mechanisms. First, deontological and altruistic decisions are more intuitively and affectively grounded, increasing their susceptibility to simple cue-based influence (J. D. Greene et al., 2004). Second, prosocial advice aligns with social desirability concerns, thereby lowering psychological resistance and the anticipated risk of social disapproval (Janoff-Bulman et al., 2009). Third, utilitarian and egoistic arguments typically require greater justification and cognitive elaboration; when such elaboration is introduced, cognitive load increases, thereby weakening the heuristic efficiency that makes advice persuasive. The pattern was especially pronounced in Study 4: participants who initially selected egoistic or utilitarian options were substantially more likely to shift toward altruistic or deontological positions when exposed to contradictory advice, whereas reverse shifts were rare. In this sense, normatively aligned advice functioned not only as a heuristic but as a social corrective, restoring coherence with internalized moral standards and reducing cognitive dissonance.

7.2.3 Advisor Irrelevance: Functional Equivalence of Human and AI Advisors

Across all four studies, AI advice exerted comparable influence to human advice. Neither in single-advisor conditions (Studies 1 and 2), nor in conditions with direct comparison (Study 3),

nor in post-decisional contexts (Study 4) did the advisor reliably affect decision outcomes. This consistent pattern of advisor irrelevance aligns with research showing that AI advice is often relied upon when it appears competent, relevant, or normatively appropriate (Krügel et al., 2025; Logg et al., 2019). When advice is brief, binary, and heuristically structured, its persuasive impact is derived primarily from its content and cognitive function rather than from its origin. Even when participants reported subjective preferences or interpersonal similarity effects, these explained only marginal variance in behavior. Identity- and affect-based cues thus played a secondary role, whereas the effort-reducing, decision-guiding function of the advice itself remained the dominant factor shaping moral decision-making.

7.2.4 Advice Timing and the Malleability of Moral Decisions

By combining pre-decisional (Studies 1–3) and post-decisional (Study 4) advisory contexts, this dissertation demonstrates that heuristic influence operates across multiple temporal stages of moral reasoning. Moral decisions, often assumed to be stable once expressed, were shown to be susceptible to revision in the presence of new advice, particularly when that advice opposed the initial decision.

Importantly, the same normative asymmetry observed in pre-decisional contexts re-emerged post-decision. Norm-congruent advice (altruistic, deontological) was readily integrated into revised decisions, whereas norm-violating advice (egoistic, utilitarian) was largely rejected. This temporal consistency indicates that the underlying cognitive mechanisms (effort reduction, social conformity, moral coherence, and intuitive processing) operate not only before a decision is made, but also during retrospective re-evaluation.

7.2.5 Cross-Study Convergence: Moral Advice as a Multi-Functional Heuristic Cue

Taken together, the results across all four studies depict moral advice as a multi-functional heuristic cue that satisfies several psychological functions simultaneously. It reduces cognitive demands, regulates affect by lowering uncertainty, promotes alignment with socially accepted norms, and resolves internal dissonance by restoring moral coherence.

Artificial advisors fit seamlessly into this structure. When their advice is normatively aligned and presented in a heuristically efficient format, they become cognitively interchangeable

with human advisors. When their advice deviates from social norms or requires elaborate justification, their persuasive impact diminishes to a certain degree.

Across diverse moral scenarios involving life-and-death decisions, interpersonal conflicts, conflicting advisory cues, and post-decisional revision, the underlying architecture of heuristic advice-taking remained remarkably stable. Collectively, the four studies therefore demonstrate not only the persuasive capacity of artificial moral advice, but also the cognitive mechanisms that render such influence likely—particularly in an increasingly digital and advice-saturated moral environment. The following chapter puts the results into perspective of established concepts in literature as well as timely scientific research.

7.3 Theoretical Integration: Moral Advice as a Heuristic and Effort-Reduction Mechanism

The convergent empirical pattern across all four studies suggests, that the persuasive power of artificial moral advice is best understood through the framework of heuristic processing and cognitive effort reduction. Drawing on the Elaboration Likelihood Model (ELM) (Petty & Cacioppo, 1986), heuristic decision-making (Frederick, 2005; Gigerenzer & Gaissmaier, 2011; Gigerenzer & Todd, 2001; Sunstein, 2005; Tversky & Kahneman, 1973, 1974), and the effort-reduction framework (Shah & Oppenheimer, 2008), we argue that AI-generated moral advice operates primarily as a low-effort shortcut rather than as a catalyst for reflective, content-based reasoning.

7.3.1 Advice as a Peripheral Cue in the Elaboration Likelihood Model

According to the Elaboration Likelihood Model (ELM), persuasion proceeds via two routes: a central, effortful route based on systematic evaluation of arguments, and a peripheral route that relies on simple cues and heuristics when motivation or cognitive ability is low (Petty & Cacioppo, 1986; Shah & Oppenheimer, 2008).

Moral dilemmas, by their very nature, are characterized by a high degree of cognitive and emotional complexity. They often involve conflicting values, the potential for harm, and the absence of a clearly "correct" solution. Individuals are required to weigh incompatible moral principles, tolerate ambiguity, and accept that any choice may involve moral cost. As a

result, moral decision-making is inherently cognitively demanding, emotionally taxing, and experienced as a form of dilemma in the literal sense, a psychological double bind in which no option is fully satisfactory (Christensen & Gomila, 2012; Duff & Sinnott-Armstrong, 1989).

Such conditions are precisely those under which motivation or cognitive capacity for deep elaboration is frequently reduced. In everyday life, individuals do not typically engage in prolonged, systematic moral reasoning. Instead, they must make decisions under time pressure, emotional arousal, distraction, or cognitive load. From the perspective of the ELM, these situational constraints reduce the likelihood of central route processing and increase reliance on the peripheral route.

Humans, as boundedly rational agents, are not designed to maximize accuracy at all costs, but to manage cognitive resources efficiently. Faced with informational overload and constant decisional demands, the human cognitive system is optimized to minimize effort whenever possible. This tendency toward effort minimization is not a deficiency but an adaptive feature that allows individuals to function in complex environments. One of the primary mechanisms through which this efficiency is achieved is the use of heuristics (Gigerenzer & Todd, 2001; H. A. Simon, 1990). Heuristics are short-cuts in judgment that reduce cognitive load by simplifying complex information, narrowing the decision space, and substituting difficult questions with more accessible ones (Gigerenzer & Gaissmaier, 2011; Tversky & Kahneman, 1973, 1974). In morally demanding contexts, this process becomes particularly salient. Instead of asking, "What is the most ethically justified action, considering all possible consequences, principles, and stakeholders?", individuals rely on simpler proxies such as "What do others recommend?" or "What seems socially acceptable?" or "Which option feels less wrong?" (Rand et al., 2014; Sunstein, 2005).

Within the framework of the ELM, advice functions as precisely such a peripheral cue. Rather than engaging in effortful evaluation of moral arguments, individuals can adopt the advice as a ready-made conclusion. The presence of advice transforms a complex moral dilemma into a binary choice: follow or reject the advice. This drastically reduces the need for internal deliberation, moral conflict resolution, and cognitive integration. In this sense, advice operates as an externally provided heuristic, being a cognitive shortcut that allows individuals to resolve complex moral tension with minimal mental effort (Krügel et al., 2023c;

Leib et al., 2023).

This interpretation is strongly supported by the empirical findings of the present dissertation. Across all four studies, participants consistently relied on advice, regardless of its advisor, level of difficulty, or justification. Even when the advice opposed their prior decision (Study 4) or when AI and human advisors provided conflicting advice (Study 3), participants frequently adjusted their moral decisions in alignment with the AI advice received. This pattern is difficult to reconcile with a pure central-route processing model. Instead, it suggests that advice is often processed peripherally, as a cue that alleviates cognitive burden rather than as an argument that is critically assessed.

Moreover, increasing the amount of argumentation did not reliably increase persuasive strength. In some cases, elaboration even reduced the impact of the advice, particularly when it promoted socially non-preferred outcomes (e.g., utilitarian harm). This aligns with the assumption that excessive information shifts processing from the peripheral route to the central route, thereby increasing cognitive effort instead of reducing it. In such cases, advice loses its heuristic function and instead becomes a stimulus for reflection, conflict, and deeper evaluation. From this perspective, effective advice does not operate as a complex moral argument but rather as a simple, intelligible, and socially anchored cue. Especially when motivation or ability is limited (which also reflects the default condition in many real-life moral situations), advice becomes an efficient tool for navigating moral complexity. It allows individuals to offload cognitive and moral effort to an external system, thereby increasing perceived decisional ease while reducing the discomfort associated with moral uncertainty.

The findings of this dissertation strongly suggest that moral advice, particularly when delivered by artificial systems, is predominantly processed through the peripheral route of the ELM. It acts as an effort-reducing heuristic that simplifies morally demanding situations by offering a clear and readily adoptable course of action. This mechanism helps explain both the strength and the potential risk of AI-based moral advice: what makes it highly effective is precisely the fact that it circumvents deep moral reflection.

Additionally, the conditions of the present studies (online participation, abstract dilemmas, complex trade-offs, and absence of real-world consequences) further make peripheral processing highly plausible. Participants often faced decisions requiring considerable cognitive effort,

emotional regulation, or moral integration. Under such constraints, externally provided advice served as a readily available cue that reduced cognitive demands.

Why Artificial Advice is Fundamentally Different to Human Advice even though their Influence is Comparable

What fundamentally distinguishes the present findings from prior research on advice-taking is the nature of the advisor. For the first time, individuals are increasingly confronted with a non-human, algorithmic source that can independently generate language, moral framing, and advice. AI systems are no longer merely passive tools but operate as autonomous, continuously available advisory agents embedded in everyday environments (e.g., smartphones, wearables, digital assistants). This development marks a structural transformation in the ecology of moral decision-making. In traditional settings, advice is episodic, socially bounded, and embedded in reciprocal interpersonal relationships. In contrast, artificial advice is scalable, omnipresent, and available without social friction, effort, or relational cost. As a result, fast and effortless assisted decision-making via the peripheral route is no longer activated occasionally but can become chronically integrated into everyday moral judgment.

From a heuristic perspective, AI therefore constitutes a qualitatively new class of peripheral cue: an always-accessible, cognitively fluent, and socially unburdensome source of direction. Crucially, its persuasive potential does not stem from superior argument quality or epistemic authority, but from its structural availability, immediacy, and perceived objectivity. In this sense, artificial advice may amplify the human tendency to minimize cognitive and moral effort, thereby increasing both its functional utility and its potential to reshape long-term patterns of moral reasoning and autonomy.

7.3.2 Why Artificial Advice Is Especially Well Suited as a Heuristic

Heuristics simplify decision-making by reducing the informational, cognitive, and motivational burdens associated with complex judgments. In the context of moral dilemmas, advice fulfills the defining features of a heuristic particularly well. According to the effort-reduction framework (Shah & Oppenheimer, 2008), heuristics work by:

1. examining fewer cues,

2. reducing memory and retrieval demands,
3. simplifying weighting principles,
4. integrating less information, and
5. considering fewer alternatives.

Artificial moral advice exhibits the majority of these characteristics. Across all studies from the present research, advice narrowed attention to a single moral direction, eliminated the need for retrieving or balancing competing principles, collapsed complex trade-offs into a binary choice, and reduced the decision space itself. Participants did not need to compute outcomes, evaluate consequences, or resolve internal conflict. Instead, they could adopt the suggested option as a cognitively efficient response. The empirical consistency of advice-following across tasks, sources, and temporal positions strongly supports this heuristic interpretation.

Why Artificial Advice is the Objectively Better Heuristic than Human Advice Artificial advisors are especially effective in fulfilling these functions and, in several important respects, may even exceed human advisors in their capacity to reduce cognitive effort.

First, AI systems are constantly available. Unlike human advisors, they do not depend on social coordination, time availability, or reciprocal interaction. They can be consulted instantly, at any moment of uncertainty. This frictionless accessibility substantially lowers the threshold for advice-seeking and strengthens the tendency to rely on external advice in cognitively demanding situations.

Second, artificial systems are able to draw on vast and heterogeneous information sources. While the user is presented with a short, simplified advice, this output is generated on the basis of large amounts of aggregated data, patterns, and prior examples. This creates the impression that the advice is both informed and comprehensive, allowing individuals to rely on a simple output while assuming that complex processing has already taken place. In this way, AI performs an externalization of cognitive work, compressing complexity into a cognitively effortless cue.

Third, artificial advisors are often perceived as more objective and consistent than humans (Giubilini & Savulescu, 2018; D. S. Kumar, 2025; B. Liu, 2021). Human advice is embedded

within personal perspectives, emotional states, and relational contexts. In contrast, artificial systems are frequently attributed a form of procedural neutrality and consistency, even if this objectivity may be partially illusory. This perceived impartiality can reduce critical resistance and enhance the willingness to adopt the advice, strengthening its functioning as a heuristic shortcut rather than a claim to be deeply evaluated.

Fourth, interacting with artificial systems involves minimal social cost. Asking for advice from another person may trigger concerns about judgment, indebtedness, or impression management. Advice from AI, in contrast, is anonymous, psychologically safe, and free from interpersonal expectations. This absence of social friction further encourages reliance on artificial advice in sensitive or morally ambiguous contexts.

Fifth, artificial advisors are becoming deeply embedded in everyday environments. Through smartphones, smart devices, and digital assistants, AI advisors accompany individuals across contexts, rather than appearing only in isolated interactions. As a result, artificial advice is not merely occasional, but increasingly forms a constant background presence that scaffolds perception and decision-making. Over time, this integration can transform advice from a deliberate consultation into a habitual and almost automatic response to uncertainty.

These characteristics suggest that artificial systems represent a qualitatively new class of heuristic cue. They are not simply substitutes for human advisors, but highly scalable, consistently available, and socially unburdened sources of simplified advice. Precisely because they reduce cognitive effort so effectively, they are uniquely positioned to shape moral and everyday decision-making in subtle, persistent, and potentially far-reaching ways.

It is important to emphasize that this does not imply that Artificial Moral Advisors are intrinsically “better” or “worse” than human advisors in a normative or ethical sense. Rather, artificial systems satisfy a greater number of structural criteria that define an effective heuristic: they reduce cognitive effort, minimize social and emotional costs, and provide highly convenient access to simplified advice. From a purely functional perspective, this makes artificial advice particularly well-suited to operate as a heuristic in morally demanding situations. Given its convenience, scalability, and increasing integration into everyday life, the growing role of artificial advice in human moral decision-making appears not only plausible, but largely inevitable.

7.3.3 Why Artificial Advice Does Not Have the Same Flaws of Classical Heuristics

Traditional heuristics are often criticized for systematically distorting judgment. By reducing information, narrowing attention, and relying on salient but often unrepresentative cues, heuristics can lead to bias, inconsistency, and suboptimal decisions (Kahneman, 2011; Tversky & Kahneman, 1974). Their effectiveness is achieved precisely through the exclusion of relevant information, making them efficient but frequently inaccurate.

Artificial advice, however, introduces a notable deviation from this classic limitation. While it is experienced by the user as a simple, effort-reducing cue, it is typically generated through the integration of vastly more information than any individual human could feasibly process. Behind a concise recommendation lies an extensive aggregation of data, probabilistic patterns, linguistic structures, and context-sensitive associations. In this way, artificial advice preserves the subjective efficiency of a heuristic while partially compensating for its traditional informational poverty.

In other words, AI systems enable a form of informational compression without informational neglect. The user is presented with a reduced, simplified output, but the underlying process is not characterized by the neglect of relevant inputs, as in classical heuristics. Instead, complexity is shifted from the individual to the system. Cognitive effort is offloaded externally, not by ignoring information, but by outsourcing its integration.

This creates a paradoxical transformation of the heuristic mechanism: artificial advice retains the hallmark characteristics of a shortcut (speed, simplicity, low cognitive demand) while simultaneously mitigating some of the very weaknesses that historically defined heuristic processing. As a result, what appears to the individual as a low-effort cue may actually be grounded in a comparatively high level of internal complexity.

Yet, this apparent advantage comes with a subtle epistemic risk. Because the process remains opaque to the user, the perceived reliability of artificial advice may exceed its actual validity. The individual no longer distinguishes between reduced complexity and relocated complexity. This perceived epistemic superiority may further encourage reliance, reduce critical reflection, and accelerate the delegation of moral judgment to external systems (Bleher & Braun, 2022).

Thus, artificial advisors do not abolish heuristics, rather, they transform them. They

preserve the functional benefits of cognitive efficiency, while altering the distribution of effort, control, and epistemic responsibility within the decision-making process itself.

7.3.4 When Argumentation Interrupts Heuristic Processing

Although advice generally functioned as a heuristic, the findings also indicate that extended justification does not necessarily enhance persuasive impact. In Study 1, utilitarian advice accompanied by detailed argumentation were less persuasive than concise deontological cues. This suggests that elaboration may disrupt heuristic processing by increasing cognitive load, drawing attention to morally aversive trade-offs, or activating central-route scrutiny.

From an ELM perspective, increased elaboration shifts users away from low-effort, cue-based processing and toward more reflective, argument-driven judgment. While such central-route processing is typically associated with higher-quality attitude change, in the context of moral dilemmas it may instead reintroduce complexity, emotional discomfort, and internal conflict. As a result, the simplicity that makes advice an effective heuristic shortcut is weakened. Rather than facilitating persuasion, argumentation may therefore undermine the very efficiency that makes advice psychologically attractive under conditions of cognitive strain.

At the same time, when central-route processing is activated, the normative content of the advice becomes even more salient. Under conditions of deeper elaboration, individuals are reminded not only of the practical implications of a choice but of their perceived moral duty and social responsibility. In this mode, advice that aligns with socially accepted, prosocial, or harm-avoidant norms gains additional persuasive force, as it resonates with internalized cultural expectations and self-concepts as a "moral" individual. In contrast, advice that deviates from these norms (such as utilitarian endorsements of instrumental harm) becomes harder to justify upon reflection, increasing resistance rather than compliance. Thus, elaboration does not simply weaken heuristic processing, but systematically shifts persuasion in favor of normatively accepted moral positions.

7.3.5 Normative Alignment as a Social Corrective

Across studies, normatively aligned advice (altruistic or deontological) exerted a stronger influence than norm-deviant advice (egoistic or utilitarian). I interpret this asymmetry as

evidence that moral advice operates not only as an effort-reduction mechanism but also as a socially corrective cue. Advice that affirms widely shared norms (e.g., helping, avoiding harm) reduces dissonance and restores moral coherence, making it easier to adopt. In contrast, norm-violating advice threatens moral identity and is readily discounted, preserving self-consistency with minimal effort.

In Study 4, this dynamic became especially visible: altruistic and deontological advice more often triggered post-decisional changes, whereas egoistic and utilitarian counter-advice was predominantly resisted. This suggests that advice does not operate neutrally but is filtered through culturally learned expectations of what one ought to do as a moral member of society. In moments of uncertainty or conflict, advice that reminds individuals of these shared moral responsibilities functions as a corrective signal, guiding them back toward socially sanctioned standards of conduct. Consequentially, moral advice operates on two interrelated levels. First, it functions as a heuristic that reduces cognitive effort in situations characterized by moral complexity and uncertainty. Second, when more elaborate processing is engaged, it activates a normative reference frame that foregrounds social obligations, moral duties, and culturally sanctioned standards. Under both processing routes, normatively aligned advice retains a systematic advantage, either because it can be adopted with minimal cognitive investment or because it is more easily justified and defended under reflective moral scrutiny.

7.3.6 Advisor Irrelevance and Heuristic Social Processing of AI

The consistent absence of advisor effects across all four studies supports the proposition that AI has become a socially legitimate provider of moral cues. As long as advice reduces cognitive effort and aligns with shared moral norms, participants appear largely indifferent to whether the advisor is human or artificial. This finding extends recent research suggesting that AI agents are increasingly processed not merely as technical tools, but as socially meaningful and persuasive sources in their own right. Crucially, this influence does not depend on anthropomorphism, interpersonal warmth, or human-like authority. Instead, it follows from the functional role of advice as a cognitively efficient cue that enables rapid resolution of moral uncertainty.

These converging perspectives point to a unified interpretation: when moral advice is concise,

normatively aligned, and unburdened by extensive justification, it operates as a peripheral-route heuristic in morally complex situations. Artificial advisors are particularly well-suited to this role because of their consistency, clarity, and ubiquitous accessibility. Consequently, AI-generated moral advice exerts persuasive power not because it delivers superior moral reasoning, but because it offers an immediately available, low-effort solution to situations characterized by internal conflict, ambiguity, or cognitive strain.

The next section turns to the broader societal and ethical consequences of increasing reliance on such effort-reducing artificial moral cues.

7.4 Societal and Ethical Practical Implications: Moral Offloading in an Age of Artificial Advisors

The conclusion that artificial moral advice functions as a superior effort-reducing heuristic has profound implications for individual moral agency, social responsibility, and society at large. As AI systems become increasingly embedded in everyday technologies, like messaging platforms, digital assistants, wearables, productivity tools or integrated operating systems, the conditions under which individuals rely on advice-based shortcuts will expand dramatically (Bonnefon et al., 2020; Cervantes et al., 2020). Proposals of Artificial Moral Advisors (AMAs) as impartial "ideal observers" that help individuals implement their own values (Giubilini & Savulescu, 2018) and of dialogical systems such as SocrAI as artificial moral experts (Rodríguez-López & Rueda, 2023) already illustrate how moral advice can be tightly integrated into everyday decision-making. The persuasive mechanisms identified across our studies therefore cannot be viewed merely as psychological curiosities. They represent emerging societal dynamics that may alter how people deliberate, justify, and assume responsibility for moral action (Purcell & Bonnefon, 2023).

7.4.1 Moral Offloading and the Erosion of Moral Practice

Heuristic reliance on external advice reduces cognitive effort in the moment but raises concerns about long-term consequences. Moral competence, like any skill, develops through repeated engagement with conflict, deliberation, and reflective self-correction. When decisions are routinely simplified through algorithmic cues, the frequency and depth of such engagement

may decline. This dynamic parallels the phenomenon of cognitive offloading in memory and problem-solving, where dependence on external tools gradually reduces internal competence (Badžak et al., 2024). Applied to moral cognition, habitual reliance on artificial advisors may diminish the capacity for autonomous moral reasoning, empathy-based evaluation, and perspective integration. Once moral practice is outsourced, the motivational and cognitive "muscles" associated with ethical reflection may atrophy.

Recent work on AI and moral corruption suggests that these concerns are not merely speculative. Köbis et al. (2021) distinguish between artificial systems that act as influencers (e.g., advisors or role models) and those that function as enablers (e.g., partners or delegates) of unethical behavior. As influencers, AI advisors may not be more corruptive than humans on a single occasion, but their ability to engage in personalized mass persuasion creates substantial aggregate influence across many interactions. As enablers, they provide a "dangerous trifecta" of opacity, anonymity, and social distance, allowing people to psychologically dissociate from questionable acts and to maintain a positive moral self-image while pursuing self-serving goals (Köbis et al., 2021, p. 13). Empirically, people follow dishonest or corruptive AI-generated advice just as readily as identical advice from humans and evaluate the wrongness of the act and the responsibility of the advisor in comparable ways (Leib et al., 2023).

Combined with the present findings on strong advice-following and reduced perceived responsibility, these results support the concern that AMAs may facilitate a mode of "moral autopilot" in which agents increasingly rely on opaque external systems to navigate ethically charged situations. Rather than fostering phronesis and the capacity to grapple with what some authors describe as intrinsically tragic, irresolvable moral conflicts (Sommaggio & Marchiori, 2020), frequent reliance on artificial advice risks encouraging a strategy of avoidance and delegation. As AI systems increasingly intervene proactively with e.g. detecting conflicts, flagging moral tension, and proposing "helpful" resolutions, the threshold for engaging in reflective moral thinking may rise, making intuitive shortcut strategies even more dominant in daily life (Etienne, 2022).

7.4.2 Responsibility Diffusion and the Reallocation of Moral Accountability

A key consequence of moral offloading is the diffusion and reallocation of responsibility. When individuals adopt externally provided advice, they may attribute part of the decision's moral weight to the advisory system rather than to themselves. This outsourcing of accountability parallels classical findings from social psychology on the bystander effect and group decision-making, where shared responsibility leads to diminished individual accountability (Fischer et al., 2011). Across our studies, advice in some cases reduced perceived responsibility. When (artificial) advice serves as a heuristic, decisions feel easier and less self-authored, encouraging individuals to interpret outcomes as partially determined by the advisor rather than as fully their own.

Debates on the moral status of AMAs sharpen this issue. Philosophical analyses generally agree that non-sentient systems cannot meet the core conditions for moral responsibility, such as originating action from within, possessing freedom in a robust sense, or engaging in self-directed deliberation (Constantinescu et al., 2022). From this perspective, AMAs are not appropriate bearers of praise or blame; moral responsibility ultimately resides with the human users, developers, and institutions that design and deploy them. Yet empirical work suggests that lay attributions of responsibility are more ambivalent. In AI-assisted bail decisions, people ascribe substantially higher degrees of forward-looking responsibility (e.g., task, power, obligation) to human agents, but they distribute backward-looking responsibility (e.g., blame, liability, causal responsibility) almost equally between humans and AI systems (Lima et al., 2021).

This pattern indicates that individuals sometimes treat AI as a quasi-responsible entity while simultaneously using it to dilute or redistribute human accountability. Studies of joint human–AI decision-making further show that when an AI system is involved, human decision-makers are often faulted less than when acting alone, and the resulting actions are sometimes judged as more permissible (Shank et al., 2019). Combined with widespread optimism about AI's fairness and objectivity, this opens the door for agency laundering: actors may strategically hide behind the system, shifting blame to the algorithm while downplaying their own role in setting goals, selecting inputs, or accepting recommendations (Y. Liu & Moore, 2025). Together with the observed readiness to follow corruptive advice from AI

(Leib et al., 2023), these dynamics create a double asymmetry. Normatively, AMAs lack the properties required for genuine moral responsibility (Constantinescu et al., 2022), yet psychologically, they function as convenient repositories for blame. As AI becomes an everyday feature of decision-support, this diffusion of responsibility risks becoming a pervasive social phenomenon, shifting moral accountability from agents to systems and fostering diminished civic agency, ethical disengagement, and the normalization of "moral autopilot" behavior in domains where personal involvement and critical reflection remain essential (Huang et al., 2023).

7.4.3 Normative Steering and the Amplification of Dominant Moral Norms

My findings consistently show that normatively aligned advice (altruistic and deontological cues) exerts disproportionate persuasive power. As AMAs scale across populations, this asymmetry implies that artificial systems may systematically reinforce culturally dominant norms while marginalizing minority or contextually atypical moral perspectives. Approaches that seek to align AI systems with shared human values by training them to approximate common moral judgments across scenarios (Hendrycks et al., 2023) may reduce the risk of obviously unacceptable advice but at the same time risk entrenching the preferences of moral majorities.

Critiques of large-scale studies such as the Moral Machine experiment highlight this problem. Using aggregated descriptive preferences as a basis for prescriptive design choices for autonomous systems involves a problematic move from "is" to "ought" and may legitimate discriminatory or unfair trade-offs simply because they enjoy broad support (LaCroix, 2022). Relatedly, work on moral dilemmas in the AI era argues that many conflicts are intrinsically tragic and irresolvable, resisting any attempt to encode a single correct solution (Sommaggio & Marchiori, 2020). Attempts to fix a unique moral rule into machine behavior may therefore obscure the depth of moral disagreement rather than resolve it.

Against this background, the asymmetrical influence of normatively aligned advice observed in the present dissertation can be interpreted as a form of normative steering. On the one hand, socially corrective cues can promote prosocial behavior and support individuals in living up to widely endorsed moral standards (Giubilini & Savulescu, 2018; Rodríguez-López &

Rueda, 2023). On the other hand, when such cues are delivered by scalable, persistent, and ubiquitous artificial systems, the cumulative effect may narrow moral diversity and entrench a homogenized moral landscape, especially if systems adapt their advice based on user behavior, feedback, or reward structures. In morally pluralistic societies, this raises the possibility of silent normative convergence driven by algorithmic feedback loops and preference aggregation rather than by democratic deliberation or individual reasoning (Bonnefon et al., 2020; Purcell & Bonnefon, 2023).

7.4.4 Potential Benefits and Risks of Artificial Moral Advisors

Despite the substantial concerns associated with moral offloading, responsibility diffusion, and normative homogenization, Artificial Moral Advisors also bring with them a number of potentially valuable functions. Under certain conditions, AMAs may enhance moral consistency, reduce ethically harmful impulsivity, and support reflective self-regulation in individuals who struggle with conflicting motives or overwhelming situational complexity. Proponents of AMAs have argued that such systems could help users implement their own moral commitments more reliably, acting as external scaffolds that counteract weakness of will, situational bias, or emotional volatility (Giubilini & Savulescu, 2018). In this sense, AMAs need not be understood primarily as replacements for moral agency but can, in principle, operate as cognitive prostheses that support individuals in realizing values they already endorse. Related proposals such as SocrAI conceptualize artificial systems not as prescriptive authorities, but as dialogical moral companions that stimulate ethical reasoning through questioning, perspective-taking, and moral reflection (Rodríguez-López & Rueda, 2023). Properly designed, such systems could help users recognize blind spots, challenge unjustified intuitions, or consider neglected stakeholders. Especially in emotionally charged or high-pressure situations, AI advisors might reduce reactive harm, highlight long-term consequences, and introduce morally relevant information that would otherwise be ignored. Empirical work on AI decision support in other domains further suggests that, when deployed carefully, artificial systems can improve consistency, reduce noise, and correct for certain types of systematic human bias (Bonnefon et al., 2020; Hendrycks et al., 2023).

At the same time, these potential benefits are tightly intertwined with the very risks

identified in the present work. The same mechanisms that allow AMAs to enhance consistency can also produce moral rigidity. The same capacity to reduce cognitive burden can weaken independent reasoning. And the same norm-alignment algorithms, intended to prevent harmful advice, may entrench dominant moral perspectives at the expense of pluralism (LaCroix, 2022; Sommaggio & Marchiori, 2020). Moreover, when AMAs are framed as moral experts, ideal observers, or ethically superior agents, users may defer too readily, confusing computational optimization with moral wisdom and responsiveness with legitimacy (Giubilini & Savulescu, 2018; Purcell & Bonnefon, 2023).

This ambivalence points to an important structural insight: AMAs are not intrinsically beneficial or harmful. Rather, their ethical valence depends on how they are designed, framed, and socially embedded. Systems that promote transparency, preserve disagreement, encourage reflection, and foreground human responsibility may function as tools of moral empowerment. In contrast, systems optimized exclusively for efficiency, compliance, or normative convergence risk contributing to moral deskilling, agency laundering, and the erosion of ethical pluralism (Huang et al., 2023; Köbis et al., 2021; Y. Liu & Moore, 2025). The same heuristic potency that makes AMAs psychologically effective therefore also makes them ethically volatile.

In light of this dual potential, the central ethical challenge is not whether AI should play a role in moral decision-making, but what kind of role it should play. A future in which artificial systems support human moral deliberation, without replacing, narrowing, or dissolving it, requires careful governance, interdisciplinary design, and ongoing empirical scrutiny. Without such safeguards, the gradual normalization of morally persuasive artificial agents may shift societies from reflective moral engagement toward automated moral compliance, a transition whose long-term consequences remain deeply uncertain.

7.4.5 The Transparency Tension: Trust, Explainability, and Heuristic Influence

The present findings suggest a nuanced relationship between transparency, persuasion, and perceived ethical safety. In the present research, extended argumentation did not reliably enhance the impact of moral advice; in some cases, detailed justifications for norm-violating advice were less persuasive than concise, norm-congruent cues. From a process perspective, such elaboration likely increases cognitive load, invites central-route scrutiny, and makes

morally aversive trade-offs more salient. By contrast, brief, binary advice can be processed more heuristically and therefore exerts stronger influence under conditions of limited motivation or capacity.

At the same time, ethical guidelines and public debates around AI emphasize transparency and explainability as central preconditions for trust, perceived legitimacy, and user autonomy. People frequently express the expectation that AI systems, especially those involved in morally relevant decisions, should be understandable, accountable, and open to scrutiny (Hendrycks et al., 2023; Huang et al., 2023). Work on legitimacy in political and ethical contexts further suggests that citizens may reject algorithmic systems, or opt out of using them, when they perceive them as opaque or as bypassing human responsibility, even if those systems could in principle improve decision quality (Bonnefon et al., 2020; Etienne, 2022). These strands of evidence point to a descriptive tension rather than to a simple design recommendation. On the one hand, transparency and explainability are normatively and socially valued: they are tied to trust, contestability, and the possibility of assigning responsibility. On the other hand, when advice operates primarily as a heuristic, it tends to be most influential precisely when it is concise, cognitively effortless, and relatively opaque in its internal workings.

In other words, the same individuals who demand transparency as a condition for trusting AI systems may, in practice, be more susceptible to untransparent advice because it better serves their need for cognitive ease in complex moral situations. This tension is further complicated by the plurality of ethical frameworks. Utilitarian, deontological, and virtue-ethical theories often yield divergent prescriptions, and translating any of them into machine-executable rules involves controversial normative choices (Huang et al., 2023; D. S. Kumar, 2025). Making these choices visible through explanation can support informed evaluation, but it may also expose disagreement and thereby weaken the intuitive appeal of the advice. The present data do not resolve how transparency ought to be balanced against heuristic effectiveness; instead, they highlight that empirical patterns of influence and normative expectations about explainability do not align perfectly. Clarifying how users understand, value, and respond to different forms of explanation therefore represents an important task for future research on AMAs and for governance frameworks that aim to reconcile trust, autonomy, and practical usability.

7.4.6 Moral Influence at Scale: Everyday Integration and Ubiquity

The implications of this research extend beyond isolated sacrificial or interpersonal dilemmas. As AI becomes embedded in everyday decision contexts like calendar conflicts, interpersonal scheduling, or automated prioritization, advice-based heuristics will shape moral patterns not just episodically but habitually. In such environments, suggested responses, reminders, and prioritization cues may be triggered without explicit user request, for instance by suggesting which obligations to honor first, which messages deserve timely replies, or how to balance personal and professional commitments. Over time, these micro-interventions can form a background layer of moral scaffolding that subtly structures attention, salience, and perceived obligation. Public attitudes toward such integration are ambivalent. On the one hand, people exhibit a robust aversion to delegating morally relevant decisions to machines and often evaluate the very act of delegation negatively, regardless of whether the outcome is beneficial for third parties (Gogoll & Uhl, 2018; Purcell & Bonnefon, 2023). On the other hand, this aversion appears to weaken under certain conditions. Perceptions of AI systems as possessing a human-like mind and being capable of thinking and feeling reduce negative evaluations of autonomous vehicles and increase trust and willingness to accept their decisions (Young & Monroe, 2019). Likewise, repeated exposure to AI in everyday contexts is associated with lower reluctance to rely on AI for morality-related decisions, even though AI agents are still often judged more blameworthy and less moral than humans for identical outcomes (Z. Zhang et al., 2022). When combined with the scalability of digital systems and their capacity for personalized mass persuasion (Köbis et al., 2021), these trends suggest a trajectory in which initial skepticism may gradually give way to normalized, and potentially unreflective, reliance on artificial advisors. Proposals that frame AMAs as moral experts or ideal observers (Giubilini & Savulescu, 2018; Rodríguez-López & Rueda, 2023) further illustrate how artificial advice can be integrated into self-understandings of moral agency, not merely as external tools but as routine companions in ethical reflection. The move from voluntary, episodic consultation to ambient, system-initiated suggestion represents a fundamental shift: when artificial systems anticipate moral tension and propose solutions before users consciously recognize a dilemma, the boundary between moral agency and system scaffolding becomes increasingly blurred.

7.4.7 Concentrated Power, Market Incentives, and the Risk of Normative Capture by Big Tech

Beyond the cognitive and interpersonal dynamics of moral offloading, a critical societal concern arises from the structural concentration of power among a small number of global technology corporations and influential individuals who design, train, and deploy advanced AI systems. Most contemporary large-scale AI models are proprietary, non–open source, trained on undisclosed data, and governed by internal, non-transparent decision-making procedures. This opacity creates a significant asymmetry of knowledge and power between developers and users, limiting public scrutiny over how values, assumptions, and normative priorities are embedded in these systems (Schneider et al., 2023; Taeihagh, 2025; Ustahaliloğlu, 2025).

While ethical AI principles such as transparency, accountability, fairness, and human oversight are widely endorsed, multiple reviews show that these ideals are rarely translated into binding, enforceable governance structures (Agbese et al., 2021; Corrêa et al., 2023; Papagiannidis et al., 2025). Most governmental and institutional responses remain in the form of voluntary guidelines, soft-law instruments, or broad normative aspirations without concrete technical implementation mechanisms (Corrêa et al., 2023). As a result, the practical regulation of AI systems often remains in the hands of the very corporations that profit from their deployment, creating a structural conflict between ethical governance and commercial incentives.

This risk is exacerbated by the competitive logic of global capitalism. In an environment of intense market pressure between dominant firms, such as Google, OpenAI, Meta, Microsoft, and emerging competitors in China, ethical safeguards may be perceived as barriers to speed, innovation, or market dominance. Several authors argue that in such contexts, performance optimization and market expansion tend to take priority over slower, more reflective ethical risk assessment (Papagiannidis et al., 2025; Schneider et al., 2023). The resulting "ethics–performance trade-off" is especially concerning when AI systems are increasingly applied in morally sensitive domains such as healthcare, governance, security, hiring, and judicial recommendation (Mutitu, 2024; Ustahaliloğlu, 2025).

At the same time, cultures differ substantially in how much influence people are willing to grant technological systems over human decision-making. Concepts such as privacy, fairness,

autonomy, and acceptable degrees of surveillance vary strongly between geopolitical regions (Corrêa et al., 2023; Hagerty & Rubinov, 2019). This raises the danger that a small number of culturally situated corporations may export their implicit value systems globally through widely deployed AI assistants and AI advisors. Rather than reflecting universal moral principles, such systems risk imposing localized moral frameworks on diverse populations (Binns, 2018; Crawford, 2021).

Even more critically, the development and strategic direction of dominant AI models is often shaped not only by institutions but also by highly influential individuals and closed executive boards. The decisions of a small and largely unaccountable group of leadership individuals, which are not necessarily known to the public, can therefore shape the moral architecture of globally deployed systems. In this sense, societies may become subtly susceptible not only to AI influence but also to the ethical visions, economic interests, ideological priorities, or psychological biases of a very small group of actors embedded within powerful organizations. This concern becomes especially salient when combined with the persuasive and heuristic nature of AI advice demonstrated in the present dissertation.

Furthermore, authors such as Kate Crawford (2021) argue that AI is not merely a digital or cognitive system, but part of a vast extractive infrastructure involving mineral mining, exploitative data labor, surveillance practices, and global inequality. The seeming neutrality of AI masks underlying material, ecological, and social costs borne disproportionately by vulnerable populations in developing regions (Crawford, 2021; Hagerty & Rubinov, 2019). These asymmetries challenge the notion that AI is merely a cognitive or technical tool, instead revealing it as an embedded political and economic structure.

Attempts to regulate these systems face an additional dilemma: insufficient regulation leaves societies vulnerable to abuse, manipulation, monopolization, and hidden normative influence, while excessive or poorly designed regulation risks stifling innovation, restricting beneficial applications, and slowing scientific progress. This tension resembles historical struggles with industrial monopolies, such as the nationalization and regulation of railways in the early and mid-20th century, where infrastructure of strategic importance outgrew purely private governance structures. AI may represent a similar threshold technology, one whose societal impact exceeds the ethical legitimacy of purely market-driven control.

Artificial Moral Advisors do not operate in a normative vacuum. They emerge from, and are shaped by, political economies, power hierarchies, cultural ideologies, and strategic interests. When combined with the strong heuristic influence demonstrated in this dissertation, alongside phenomena such as agency laundering, responsibility diffusion, and moral offloading the concentration of AI control in the hands of a small number of private actors poses a structural risk of large-scale, invisible, and persistent moral steering (Agbese et al., 2021; Corrêa et al., 2023; Taeihagh, 2025). This risk does not lie in malevolence alone, but in the quiet normalization of AI influence integrated into everyday life without transparent democratic negotiation.

Consequently, the ethical governance of AI must extend beyond abstract principles and technical safeguards. It requires institutional pluralism, independent oversight, cultural representation, open scientific participation, and mechanisms that prevent the consolidation of moral authority in opaque corporate structures. Without such measures, the same systems that offer unprecedented cognitive assistance may become instruments of silent normative concentration in the digital moral landscape.

7.4.8 Societal Risks and the Need for Moral Infrastructure

Ultimately, the societal impact of widespread AI advisory systems depends not only on individual cognitive dynamics but also on the institutional and regulatory infrastructure through which AI systems are integrated into daily life. Without guardrails, artificial systems may inadvertently shape public morality through accumulative nudging, asymmetric norm reinforcement, and the reduction of reflective engagement. Accountability challenges such as the "problem of many hands" (where undesirable outcomes can be traced to multiple human contributors across the AI lifecycle) complicate efforts to assign responsibility and design effective oversight mechanisms (Huang et al., 2023).

Several lines of work converge on the need for explicit moral infrastructure around AMAs. Proposals for value alignment emphasize that AI systems should be constrained by widely shared human values, while acknowledging the difficulty of operationalizing such values without marginalizing minority perspectives (Hendrycks et al., 2023). Calls for participatory and democratic governance stress that citizens must be involved in setting the ethical parameters

of machines to avoid ethically problematic forms of opt-out or backlash (Bonnefon et al., 2020; D. S. Kumar, 2025). At the same time, empirical findings on responsibility attributions and agency laundering highlight the necessity of institutional arrangements that prevent designers, deployers, and users from hiding behind algorithmic opacity when moral failures occur (Lima et al., 2021; Y. Liu & Moore, 2025; Shank et al., 2019).

In addition to regulatory oversight, there is a need for cultural and educational responses. Public literacy about cognitive offloading, moral deskilling, and the persuasive design of AI systems can help individuals recognize when they are outsourcing moral work and what trade-offs this entails (Etienne, 2022; Köbis et al., 2021). More broadly, AMAs should be embedded in environments that encourage reflective engagement rather than passive compliance, for example by offering optional explanations, surfacing value conflicts, or prompting users to consider alternative courses of action when stakes are high (Rodríguez-López & Rueda, 2023). Maybe even removing the recommendation for action and therefore just providing reflective enhancing thoughts on all alternatives.

Such measures are essential to ensure that AI advisory systems enhance rather than erode moral agency. If carefully governed, AMAs may support individuals in acting consistently with their values and in navigating complex moral environments. If left unchecked, however, they risk contributing to moral deskilling, responsibility diffusion, and the quiet consolidation of dominant norms under the guise of cognitive efficiency.

7.5 Strength and Limitations

The present dissertation employed a multi-study, experimental approach to examine the influence of human and artificial moral advice under varying conditions of dilemma type, advisory structure, and decision timing. Across studies, comparable recruitment channels, ethical standards, measurement strategies, and statistical modelling techniques were applied. This methodological consistency allows for meaningful cross-study integration, while systematic design variation enhances both internal validity and conceptual breadth. In the following, central methodological strengths and limitations are discussed.

7.5.1 Strengths

Several methodological features strengthen the internal validity, analytical rigor, and interpretability of the present research program.

Large Samples and Adequate Statistical Power All studies were based on substantial sample sizes. Study 1 included over 1,100 participants, and the remaining studies included between 300 and 400 participants per study, yielding a total of $N = 2{,}579$ across all four studies. Although repeated participation across studies cannot be fully ruled out, post-hoc power analyses indicated excellent statistical sensitivity, with achieved power values exceeding $(1 - \beta) = .99$ for all key effects. This reduces the likelihood of Type-II errors and increases confidence in the robustness and stability of the observed effects.

Experimental Control and Randomization All advice manipulations (advisor, content, congruence) were randomly assigned, both between and within participants. In Studies 1 and 2, participants were randomly assigned to a consistent advisor condition, while advice content varied across dilemmas. In Studies 3 and 4, advice content, advisor, and congruence were randomized within participants. These procedures substantially reduce potential confounds and strengthen causal inference.

Use of Established and Validated Paradigms Dilemma stimuli were primarily based on well-established moral psychology paradigms (J. D. Greene et al., 2004, 2008; Singer et al., 2019). By embedding novel research questions within empirically validated materials, the studies ensure strong construct validity and anchor the present findings in a well-replicated literature. The inclusion of everyday moral dilemmas and relational contexts extends classical paradigms in a conceptually meaningful way. The exclusion of highly politicized or socially sensitive contemporary topics (e.g., COVID-19, war) further minimized unwanted bias from pre-existing attitudes.

High Analytical Sophistication Across studies, data were analyzed using generalized linear and linear mixed-effects modelling (GLMMs and LMMs) that accounted for the hierarchical structure of the data (decisions nested within participants and dilemmas). This approach is

theoretically appropriate given that moral decisions are embedded in both individual and situational contexts, and it is statistically superior to traditional ANOVA techniques because it more effectively handles unbalanced data, missing values, and random effects (Hilbert et al., 2019). Random intercepts for participants and, where applicable, dilemmas increased the precision and ecological plausibility of the statistical models.

Binary Response Format as a Methodological Advantage Although moral cognition is inherently complex and multidimensional, the use of a binary response format (e.g., altruistic vs. egoistic; deontological vs. utilitarian) constitutes a deliberate and theoretically grounded methodological strength in the present research. Moral dilemmas, by definition, are structured around an irreducible tension between two competing and mutually incompatible courses of action. The forced-choice format therefore does not oversimplify the moral problem, but rather making the underlying conflict maximally explicit. By constraining participants to commit to one of two normatively distinct options, the paradigm accentuates the very tension that characterizes moral dilemmas and aligns closely with classical designs in moral psychology. This structural clarity increases measurement reliability, minimizes interpretational ambiguity, and enables a precise operationalization of advice adherence and deviation. In a research context specifically concerned with the directional influence of advice, such decisional constraint is essential: only when individuals are required to resolve the conflict through an explicit choice can shifts in decision and external advice-following be validly and sensitively quantified.

Programmatic Consistency with Conceptual Variation While maintaining a stable methodological backbone (online experiments, controlled advice manipulation, validated trait measures), the studies systematically varied key theoretical dimensions: (1) type and difficulty of dilemma (non-moral, impersonal moral, personal/sacrificial moral, and everyday moral conflicts), (2) social proximity to the affected target (close vs. distant others), (3) advisory structure (single advisor vs. directly conflicting advice from two different advisors), and (4) timing of influence (pre-decisional vs. post-decisional).

In particular, the inclusion of conditions in which participants received conflicting advice from a human and an artificial advisor, as well as the systematic manipulation of social proximity and dilemma type, allowed for a fine-grained investigation of how advisor, relational

context, and moral intensity interact in shaping artificial advice-following. This structured and progressive variation strengthens the claim that the observed effects reflect general psychological mechanisms rather than context-specific artefacts tied to a single dilemma type, level of social distance, source configuration, or decision phase.

Within-person Experimental Designs Especially in Studies 3 and 4, the use of within-subject manipulations of advice content, advisor and congruence increased statistical sensitivity and controlled for inter-individual differences in baseline morality, trust, and decision style. This design choice strengthens causal inference and allows for a more fine-grained analysis of how the same individual responds to varying advisory conditions.

Use of Validated Psychological Instruments Trait measures such as the Moral Identity Scale (Aquino & Reed, 2002), the SD-WISE (Thomas et al., 2022), and the AI Acceptance / Aversion scale (Sindermann et al., 2021) demonstrated high internal consistency and provided theoretically meaningful covariates. Their placement after the main decision task minimized the risk of priming effects on moral decisions.

Open Science Practices All studies were preregistered on the Open Science Framework (https://osf.io/qd95s/overview?view_only=52431cb0a45d45b1a081e87f4c356a05). Transparent reporting of analytic decisions (e.g., model structures, convergence constraints, coding choices) increases the reproducibility and credibility of the findings and aligns with current best practices in psychological science.

7.5.2 Limitations

Despite the substantial methodological strengths of the present research program, several limitations constrain the generalizability of the findings and must be considered when interpreting the results.

Hypothetical and Consequence-free Design All moral decisions were made in hypothetical, consequence-free online settings. While this affords a high degree of experimental control, it necessarily reduces emotional intensity, embodied engagement, and real-world stakes.

Participants may therefore respond more reflectively, less affectively, or in a more socially desirable manner than they would in situations involving tangible interpersonal consequences, financial costs, or real harm. As a result, the absolute magnitude of advice effects may differ in ecologically valid contexts.

Limited Ecological Realism of Advisory Interaction Advice was presented in a static, text-based, and highly standardized format, which lacks many characteristics of real-world advisory interactions, such as tone of voice, timing, non-verbal cues, emotional expression, and dialogical exchange. Even in studies where visual cues of a human advisor were introduced, the interaction remained artificial and non-interactive. Consequently, the present design may underestimate the complexity and persuasive dynamics of real-world human–AI advisory relationships.

Absence of Manipulation Checks for Advisor Perception in Early Studies In Studies 1 and 2, no direct manipulation check was included to verify whether participants consciously distinguished between human and artificial advisors. It therefore remains unclear whether the observed advisor irrelevance reflects genuine cognitive equivalence or partial neglect of the manipulation. Although this limitation was addressed in Studies 3 and 4 through the inclusion of explicit manipulation checks, it constrains the interpretability of the earliest findings.

Static and Non-personalized Advice Across all studies, advice was scripted and identical for participants within the same condition. While this ensured strong experimental control, it diverges from the way in which contemporary AI systems increasingly operate: adaptive, personalized, and context-sensitive. Because non-personalized, one-shot advice may be perceived as less relevant, credible, or engaging, the present findings may underestimate the persuasive potential of future Artificial Moral Advisors.

Timing of Measurement and Potential Rationalization Effects In Studies 1–3, subjective evaluations of advice (e.g., helpfulness, certainty, perceived responsibility) were assessed after participants had already made their decision. This introduces the possibility of post-hoc rationalization, whereby participants retrospectively interpret advice as helpful simply because

they followed it, rather than because it causally influenced their choice. Although Study 4 partially addressed this by focusing on post-decisional influence, the timing of evaluation remains a methodological constraint.

Sample Homogeneity and Cultural Context Participants were primarily drawn from Western, academically affiliated populations, with a strong dominance of German students or individuals familiar with experimental research settings and AI-related discourse. Cultural differences in moral reasoning, authority perception, and trust in technology therefore limit the generalizability of the findings to non-Western, less educated, or culturally diverse populations.

Indirect Assessment of Cognitive Load Although the theoretical framework emphasizes advice as an effort-reducing heuristic, the studies did not directly manipulate or measure cognitive load, attentional resources, time pressure, or emotional arousal. As a result, the assumption that advice-following increases specifically under conditions of resource depletion remains inferential rather than empirically demonstrated within the present data.

Potential Influence of Social Desirability The stronger effect of altruistic and deontological advice may partially reflect social desirability biases rather than solely genuine moral persuasion. Participants might have felt motivated to select socially approved options or avoid endorsing norm-violating behavior, which could have amplified the asymmetric influence of normatively desirable advice.

7.5.3 Methodological Implications for Interpretation

Taken together, the methodological profile of this dissertation suggests that the observed effects are internally robust and theoretically and practically informative, yet should be interpreted with caution in terms of ecological generalizability. The controlled and minimalistic advisory setting likely underestimates, rather than overestimates, the persuasive capacity of future AI systems embedded in dynamic, multimodal, and personalized environments.

Importantly, these limitations do not undermine the central conclusion that artificial advice functions as a powerful heuristic in moral decision-making. Rather, they highlight the need for

further research under increasingly realistic conditions to examine how this heuristic operates as AI becomes more deeply integrated into everyday life.

7.6 Future Research Directions

Building on the present findings, several avenues for future research emerge that are essential for clarifying the mechanisms, boundary conditions, and long-term societal implications of AI-generated moral advice. Because the current results consistently suggest that individuals treat artificial moral advice as a cognitive shortcut, future work must examine when, why, and with what consequences people rely on such heuristics. The following directions aim to advance theoretical precision, ecological realism, and ethical understanding.

Direct Tests of the Heuristic Hypothesis: Cognitive Load and Time Pressure A central claim emerging from the present dissertation is that artificial moral advice functions as an efficient heuristic that reduces cognitive effort. However, none of the studies directly manipulated cognitive resources. Future research should therefore experimentally examine whether advice-taking increases under conditions of constrained cognitive capacity, such as when participants are placed under cognitive load through concurrent memory or attention tasks, when decisions must be made under severe time pressure, when individuals experience fatigue or ego depletion, or when they are confronted with information overload.

According to the Elaboration Likelihood Model (Petty & Cacioppo, 1986), individuals shift toward peripheral-route processing when either cognitive ability or motivational capacity is reduced. Under such conditions, simple external cues, like artificial advice, should gain disproportionate influence. Systematically varying these parameters would provide a stringent experimental test of the heuristic mechanism proposed in this dissertation and clarify whether AI advice is particularly powerful when individuals lack resources for careful moral reasoning.

Transparency, Explanation Depth, and their Paradoxical Effects The present findings suggest that increased explanation or argumentative elaboration does not necessarily strengthen the persuasive impact of advice; in some contexts, it appears to reduce it by increasing cognitive load and activating reflective scrutiny. This raises a critical empirical question: does increased

transparency reduce the heuristic appeal of advice by shifting processing from a peripheral to a more central route? Future studies should therefore systematically vary the length, complexity, and explanatory depth of advice, manipulate whether advice is presented with or without explicit justificatory arguments, and examine whether explanations emphasize socially desirable or counter-normative outcomes. Investigating this would clarify when transparency and explainability strengthen trust and reflective engagement and when they paradoxically undermine persuasive impact by removing the low-effort quality that makes advice influential in the first place. This question is especially relevant in the context of current debates on explainable and ethical AI.

Personalization versus Erosion of Moral Independence As AI systems increasingly tailor recommendations to individual users, an important psychological and ethical tension emerges between improved alignment and reduced autonomy. Future research should explore whether personalization increases acceptance and perceived relevance while still preserving independent moral reflection, or whether repeated exposure to individualized advice gradually erodes autonomous moral reasoning. Longitudinal designs are particularly important here. They would allow researchers to determine whether personalized advice leads to a form of "moral overfitting," in which individuals increasingly defer to AI advice and engage less with internal reflection, similar to the deskilling and cognitive offloading effects observed in domains such as navigation, calculation, and memory.

Normatively Favored versus Norm-deviant Advice One of the most consistent findings across the present studies was the asymmetry between normatively favored advice (altruistic or deontological), and norm-deviant advice (egoistic or utilitarian). Future research should directly examine the extent to which this asymmetry is driven by heuristic efficiency versus normative conformity. This could be achieved by systematically manipulating whether advice aligns with or violates dominant social and moral norms, by varying the degree of norm deviance, and by embedding advice in different cultural and moral contexts. Such work would clarify whether artificial advisors primarily function as neutral cognitive shortcuts or whether they amplify existing social norms and moral biases when repeatedly presenting norm-congruent advice.

Comparison with Advice from Real Peers and Social Networks To increase ecological validity, future studies should compare artificial advice not only with generic or abstract "human" advice but also with advice provided by real peers, friends, in-group members, out-group members, and trusted authority figures. In everyday life, moral advice rarely comes from anonymous, decontextualized advisors; rather, it is embedded in social relationships and networks. Comparing the persuasive impact of AI-generated advice with that of actual peers and social referents would allow researchers to situate artificial advisors within the broader landscape of social influence.

Real-world Behavioral Paradigms and Consequential Moral Choices To strengthen ecological validity, future research should employ paradigms that involve real, tangible consequences for participants or others. These could include economic trade-offs with real beneficiaries, honesty and rule-breaking tasks involving incentives, interpersonal helping scenarios, charitable allocation tasks, or immersive moral choices in virtual or augmented reality environments. Such paradigms allow researchers to test whether artificial advice influences not only stated intentions or decisions but actual morally relevant behavior in contexts characterized by accountability, emotional engagement, and potential social evaluation.

Longitudinal and Developmental Trajectories of Moral Outsourcing A central unanswered question concerns the long-term effects of repeated reliance on moral AI advice. Longitudinal and repeated-exposure studies are needed to determine whether people develop stronger habits of advice-taking over time, whether confidence in one's own moral reasoning declines, whether individuals begin to passively anticipate AI advice instead of engaging in active reflection, and whether moral intuitions themselves shift as a result of repeated AI framing. Future research could examine whether repeated reliance on AMAs is associated with progressive "moral deskilling" and a gradual shift in perceived moral responsibility.

Moral Advice in Social, Institutional, and Collective Contexts Moral decision-making is rarely confined to isolated individuals; it often takes place in social and institutional environments. Future research should therefore investigate how artificial moral advice operates in dyadic relationships, group decision-making settings such as committees or teams, and

institutional contexts such as healthcare triage, personnel selection, judicial recommendations, and policy advising. This would make it possible to examine how AI advice interacts with conformity pressures, authority hierarchies, collective responsibility, and institutional accountability structures, thereby moving closer to the real contexts in which moral AI advice is likely to be deployed.

Cross-cultural and Demographic Generalizability Given substantial cross-cultural variation in moral frameworks, authority relations, and attitudes toward technology, future research must expand beyond Western student samples. Studies involving non-Western populations, older adults, individuals with different religious and philosophical backgrounds, and groups with varying levels of technological literacy are essential for determining whether the heuristic use of moral advice is culturally universal or culturally contingent.

This would provide critical insight into the societal conditions under which artificial advice is accepted, resisted, or transformed.

Interactive and Dynamic Advisory Systems Finally, real-world AI advisors increasingly operate as interactive, conversational systems rather than static information providers. Future research may therefore explore multi-turn interactions with artificial advisors, including scenarios in which the system justifies its recommendations, responds to objections, negotiates alternatives, or adapts to user resistance.

These designs could illuminate how influence evolves over time, whether users become more confident or more dependent, and how individuals psychologically negotiate moral disagreement with artificial advisors in ongoing dialogical exchanges.

7.6.1 Conclusion

The integration of AI systems into everyday decision environments makes it increasingly important to determine when and how artificial advice operates as a heuristic at a fine-grained & systematic level, and which long-term cognitive, moral, and societal consequences follow from such reliance. Artificial Moral Advisors carry substantial potential to support individuals in navigating complex and demanding situations, yet they also pose significant risks as outlined previously. The research avenues outlined here represent only a limited selection of urgently

needed directions within a much broader applied research agenda. As AI continues to permeate private, social, and institutional spheres, establishing detailed, evidence-based understanding of its influence on moral decision-making will be essential for the responsible design, regulation, and integration of artificial advisory systems into society.

7.7 Final Remarks

Across four empirical studies, this dissertation examined how human and artificial advisors influence moral decisions across diverse contexts: classical sacrificial dilemmas, everyday moral conflicts, conflicting advisory settings, and post-decisional moral evaluations. Despite substantial variation in design, dilemma type, framing, and temporal positioning of advice, the results reveal a consistent and theoretically meaningful pattern: moral advice, including advice generated by AI, exerts a robust and reliable influence on human moral decision-making.

A central contribution of this work lies in identifying why such influence occurs. The cumulative evidence supports the interpretation that individuals process moral advice predominantly as a cognitive heuristic, low-effort cue that simplifies complex moral trade-offs. Participants followed advice not because of their advisor, argumentative depth, or perceived authority, but because advice offered an efficient mental shortcut: it reduced conflict, narrowed the choice space, and provided an immediate resolution in situations characterized by uncertainty, cognitive strain, or limited motivation to engage in effortful reasoning. This pattern aligns with the effort-reduction framework (Shah & Oppenheimer, 2008) and the peripheral-route assumptions of the Elaboration Likelihood Model (Petty & Cacioppo, 1986), suggesting that in many moral contexts, individuals prioritize cognitive efficiency over systematic evaluation.

This heuristic interpretation also helps explain several asymmetries observed across studies. Advice promoting altruistic or deontological positions proved especially persuasive, functioning as a form of social corrective that restored moral coherence when participants had initially favored egoistic or utilitarian options. Conversely, advice that violated social norms was readily dismissed, not because it originated from an artificial system, but because it imposed cognitive and social costs that contradicted intuitive moral expectations. In this sense, the persuasiveness of moral advice is shaped less by the advisor's identity and more by the normative and emotional structure of the moral domain itself.

The present research further demonstrates that moral AI advice can match human advice in persuasive strength across all tested contexts. Whether presented before or after the initial decision, and whether convergent or conflicting to human advice, AI advice influenced decision-making with an effectiveness indistinguishable from that of human advisors. This finding substantially extends prior work on AI advice-taking by demonstrating that AI does not merely shape factual or instrumental judgments, but can significantly affect moral decisions, an area traditionally associated with human autonomy, experience, and character.

These findings carry important societal implications. As artificial systems become deeply integrated into everyday digital environments, moral advice will increasingly be encountered not only as an explicit request, but also as an ambient feature of decision-making, embedded in scheduling tools, communication platforms, and intelligent assistants. Because such advice is processed heuristically, users may adopt moral advice with limited deliberative scrutiny, potentially outsourcing ethical reflection and weakening long-term moral agency. The risk is not simply misplaced trust in technology, but a gradual shift toward habitual reliance on external moral cues, accompanied by diffusion of responsibility and possible erosion of moral reasoning skills. At the same time, the strong persuasive capacity of norm-congruent advice highlights the potential of carefully designed systems to support prosocial behavior and moral learning (provided that such interventions remain transparent, contestable, and sensitive to moral pluralism).

If Artificial Intelligence is treated primarily as an enabler, rather than as a decision-maker that replaces human judgment, its role in moral contexts may shift in a fundamentally different direction. Instead of corroding moral agency, it could be used to foster additional moral elaboration, support reflection, and help individuals engage more deeply with their own values. In this sense, the future impact of Artificial Moral Advisors may depend less on the technology itself than on the role we choose to assign to it in our moral and social lives.

7.8 Core Contributions of the Present Dissertation

In summary, this research makes five central contributions to the study of moral decision-making and human–AI interaction. First, it provides the first multi-study, empirical demonstration that artificial moral advice functions as an effort-reducing heuristic in moral decision-

making, extending classic heuristic and dual-process models to AI advisors. Second, it identifies a robust normative asymmetry in advice-taking, showing that altruistic and deontological advice consistently outperform egoistic and utilitarian ones in persuasive power, thereby empirically linking AI influence to established theories of moral intuition and social normativity. Third, it offers systematic evidence that AI advice can reverse already-formed moral decisions, demonstrating that artificial systems can shape not only pre-decisional reasoning but also post-decisional moral revision. Fourth, it challenges widespread assumptions about human moral authority by showing that human advice does not dominate artificial advice, even under direct conflict. Finally, by integrating these empirical results with research on cognitive offloading, agency laundering, and responsibility diffusion, the dissertation introduces and substantiates the concept of moral offloading through AI as a psychologically grounded and societally relevant phenomenon, thereby bridging moral psychology, behavioral decision-making, and AI ethics in a novel and theoretically coherent way.

My findings offer a comprehensive, empirically grounded account of how and why AI-generated advice can shape human moral decision-making. They suggest that moral persuasion in the age of intelligent systems does not stem from superior reasoning, technological mystique, or anthropomorphic cues, but from a basic cognitive tendency to conserve effort in the face of complexity. As artificial advisors increasingly populate the moral landscape, understanding this mechanism becomes indispensable. Future research and governance must therefore balance the benefits of accessible moral support with the imperative to preserve moral autonomy, cultivate reflective engagement, and ensure that human ethical agency is supplemented, not supplanted by artificial systems.

As artificial systems increasingly participate in morally relevant decisions, the central ethical challenge is not whether machines can reason morally, but whether humans remain attentive to how their own moral judgments are shaped in interaction with them. This dissertation demonstrates that moral influence can operate subtly, even after individuals believe their decisions to be settled. Preserving moral autonomy in the age of AI advice does not require rejecting such advice, but sustained reflective vigilance toward the conditions under which moral responsibility is negotiated, delegated, or silently transformed. In an era of AI advice, moral responsibility is not outsourced, it is redistributed, and must be consciously reclaimed.

References

Agbese, M., Alanen, H.-K., Antikainen, J., Halme, E., Isomäki, H., Jantunen, M., Kemell, K.-K., Rousi, R., Vainio-Pekka, H., & Vakkuri, V. (2021). Governance of ethical and trustworthy AI systems: Research gaps in the ECCOLA method. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 224–229. https://doi.org/10.1109/REW53955.2021.00042

Al-Dajani, N., & Czyz, E. K. (2022). Suicidal desire in adolescents: An examination of the interpersonal psychological theory using daily diaries. *Journal of Clinical Child & Adolescent Psychology*, 1–15. https://doi.org/10.1080/15374416.2022.2051525

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? performance, cognition, and neurophysiology. *Computers in Human Behavior*, *89*, 279–288. https://doi.org/10.1016/j.chb.2018.07.026

Anthropic. (2023). Claude: An AI assistant by anthropic. https://www.anthropic.com/claude

Aquino, K., Freeman, D., Reed, A., Lim, V. K. G., & Felps, W. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology*, *97*(1), 123–141. https://doi.org/10.1037/a0015406

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Ardelt, M. (2004). Wisdom as expert knowledge system: A critical review of a contemporary operationalization of an ancient concept. *Human Development*, *47*(5), 257–285. https://doi.org/10.1159/000079154

Aristoteles, Crisp, R., & Aristoteles. (2011). *Nicomachean ethics* (15. print). Cambridge Univ. Press.

Ashoori, M., & Weisz, J. D. (2019, December 5). In AI we trust? factors that influence trustworthiness of AI-infused decision-making processes. https://doi.org/10.48550/arXiv.1912.02675

Ayala, F. J. (2010). The difference of being human: Morality. *Proceedings of the National Academy of Sciences*, *107*, 9015–9022. https://doi.org/10.1073/pnas.0914616107

Badžak, J., Đerke, F., Bašić, S., & Demarin, V. (2024). Digital dementia and cognitive decline in the era of smart gadgets. *Rad Hrvatske akademije znanosti i umjetnosti. Medicinske znanosti*, *68-69*, 50–54. https://doi.org/10.21857/ygjwrc27ky

Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., Adamkovic, M., Adamus, S., Albalooshi, S., Albayrak-Aydemir, N., Alfian, I. N., Alper, S., Alvarez-Solas, S., Alves, S. G., Amaya, S., Andresen, P. K., Anjum, G., Ansari, D., Arriaga, P., … Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in eastern, southern and western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*(6), 880–895. https://doi.org/10.1038/s41562-022-01319-5

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2023). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, *42*(28), 24516–24541. https://doi.org/10.1007/s12144-022-03573-2

Baines, J. I., Dalal, R. S., Ponce, L. P., & Tsai, H.-C. (2024). Advice from artificial intelligence: A review and practical implications. *Frontiers in Psychology*, *15*, 1390182. https://doi.org/10.3389/fpsyg.2024.1390182

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal*, *8*(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Baltes, P. B., & Staudinger, U. M. (2000). Wisdom: A metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. *American Psychologist*, *55*(1), 122–136. https://doi.org/10.1037/0003-066X.55.1.122

Batson, C. D. (2011). *Altruism in humans* (1st ed). Oxford University Press, Incorporated.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525–537. https://doi.org/10.1037/0022-3514.77.3.525

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Becker, F., Skirzyński, J., van Opheusden, B., & Lieder, F. (2022). Boosting human decision-making with AI-generated decision aids. *Computational Brain & Behavior*, *5*(4), 467–490. https://doi.org/10.1007/s42113-022-00149-y

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Berman, G., Chubb, J., & Williams, K. (2023). The use of artificial intelligence in science, technology, engineering, and medicine.

Bhuvana, M. L., Pavithra, M. B., & Suresha, D. S. (2021). Altruism, an attitude of unselfish concern for others – an analytical cross sectional study among the medical and engineering students in bangalore. *Journal of Family Medicine and Primary Care*, *10*(2), 706–711. https://doi.org/10.4103/jfmpc.jfmpc_834_20

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159. Retrieved October 24, 2025, from https://proceedings.mlr.press/v81/binns18a.html

Blasi, A. (1983). Moral cognition and moral action: A theoretical perspective. *Developmental Review*, *3*(2), 178–210. https://doi.org/10.1016/0273-2297(83)90029-1

Bleher, H., & Braun, M. (2022). Diffused responsibility: Attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, *2*(4), 747–761. https://doi.org/10.1007/s43681-022-00135-x

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports*, *11*(1), 8028. https://doi.org/10.1038/s41598-021-87480-9

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, *107*(3), 502–504. https://doi.org/10.1109/JPROC.2019.2897447

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2020, September 17). *The moral psychology of AI and the ethical opt-out problem*.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, *29*(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Byrne, D. (1997). An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships*, *14*(3), 417–431. https://doi.org/10.1177/0265407597143008

Caluori, L. (2024). Hey alexa, why are you called intelligent? an empirical investigation on definitions of AI. *AI & SOCIETY*, *39*(4), 1905–1919. https://doi.org/10.1007/s00146-023-01643-y

Carpendale, J. I. M. (2009). Piaget's theory of moral development. In U. Müller, J. I. M. Carpendale, & L. Smith (Eds.), *The cambridge companion to piaget* (pp. 270–286). Cambridge University Press. https://doi.org/10.1017/CCOL9780521898584.012

Carpendale, J. I. (2000). Kohlberg and piaget on stages and moral reasoning. *Developmental Review*, *20*(2), 181–205. https://doi.org/10.1006/drev.1999.0500

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501–532. https://doi.org/10.1007/s11948-019-00151-x

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Christensen, J., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Coleman, C., Neuman, W. R., Dasdan, A., Ali, S., & Shah, M. (2025, April 27). The convergent ethics of AI? analyzing moral foundation priorities in large language models with a multi-framework approach. https://doi.org/10.48550/arXiv.2504.19255

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Constantinescu, M., Vică, C., Uszkai, R., & Voinea, C. (2022). Blame it on the AI? on the moral responsibility of artificial moral advisors. *Philosophy & Technology*, *35*(2), 35. https://doi.org/10.1007/s13347-022-00529-z

Corrêa, N. K., Galvão, C., Santos, J. W., Pino, C. D., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & Oliveira, N. d. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K. D., MI in Healthcare Workshop Working Group, Beck, T., Collier, E., Colvis, C., Gersing, K., Gordon, V., Jensen, R., Shabestari, B., & Southall, N. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *npj Digital Medicine*, *3*(1), 47. https://doi.org/10.1038/s41746-020-0254-2

Dawkins, R. (1976). *The selfish gene* [Publication Title: The Selfish Gene]. Oxford University Press, Oxford, UK.

Decety, J., & Lamm, C. (2006). Human empathy through the lens of social neuroscience. *The Scientific World JOURNAL*, *6*, 1146–1163. https://doi.org/10.1100/tsw.2006.221

del Campo, C., Pauser, S., Steiner, E., & Vetschera, R. (2016). Decision making styles and the use of heuristics in decision making. *Journal of Business Economics*, *86*(4), 389–412. https://doi.org/10.1007/s11573-016-0811-y

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Chapter 10 motivated moral reasoning. In *Psychology of learning and motivation* (pp. 307–338, Vol. 50). Elsevier. https://doi.org/10.1016/S0079-7421(08)00410-6

Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, *151*, 109547. https://doi.org/10.1016/j.paid.2019.109547

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*(2), 245–260. https://doi.org/10.1023/A:1019637632584

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Duff, R. A., & Sinnott-Armstrong, W. (1989). Moral dilemmas. *The Philosophical Quarterly*, *39*(155), 240. https://doi.org/10.2307/2219643

Earp, B. D., McLoughlin, K. L., Monrad, J. T., Clark, M. S., & Crockett, M. J. (2021). How social relationships shape moral wrongness judgments. *Nature Communications*, *12*(1), 5776. https://doi.org/10.1038/s41467-021-26067-4

Ellemers, N., Van Der Toorn, J., Paunov, Y., & Van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Eres, R., Louis, W. R., & Molenberghs, P. (2018). Common and distinct neural networks involved in fMRI studies investigating morality: An ALE meta-analysis. *Social Neuroscience*, *13*(4), 384–398. https://doi.org/10.1080/17470919.2017.1357657

Eriksson, K., Karlsson, S., Vartanova, I., & Strimling, P. (2025, October 6). What makes AI applications acceptable or unacceptable? a predictive moral framework. https://doi.org/10.48550/arXiv.2508.19317

Etienne, H. (2022). Solving moral dilemmas with AI: How it helps us address the social implications of the covid-19 crisis and enhance human responsibility to tackle meta-dilemmas. *Law, Innovation and Technology*, *14*(2), 305–324. https://doi.org/10.1080/17579961.2022.2113669

European Commission. (2019). *A definition of AI: Main capabilities and scientific disciplines* (Technical Report). European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341

European Court of Auditors. (2024). *EU artificial intelligence ambition: Stronger governance and increased, more focused investment essential going forward* (Special Report 08/2024). European Court of Auditors. Luxembourg. https://www.eca.europa.eu/ECAPublications/SR-2024-08/SR-2024-08_EN.pdf

Evans, J., & Frankish, K. (Eds.). (2009). *In two minds: Dual processes and beyond*. Oxford University Press.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Evans, J. S. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Fabre, E. F., Mouratille, D., Bonnemains, V., Palmiotti, G. P., & Causse, M. (2024). Making moral decisions with artificial agents as advisors. a fNIRS study. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100096. https://doi.org/10.1016/j.chbah.2024.100096

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441. https://doi.org/10.1016/j.cognition.2012.02.001

Festinger, L. (1962). Cognitive dissonance. *SCIENTIFIC AMERICAN*.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517–537. https://doi.org/10.1037/a0023304

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Formosa, P., & Ryan, M. (2021). Making moral machines: Why we need artificial moral agents. *AI & SOCIETY*, *36*(3), 839–851. https://doi.org/10.1007/s00146-020-01089-6

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42. http://www.jstor.org/stable/4134953

Garrigan, B., Adlam, A. L., & Langdon, P. E. (2016). The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response

decision judgements. *Brain and Cognition*, *108*, 88–97. https://doi.org/10.1016/j.bandc.2016.07.007

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T. K., Hudecek, M. F. C., Ackery, A. D., Grover, S. C., Coughlin, J. F., Frey, D., Kitamura, F. C., Ghassemi, M., & Colak, E. (2023). Non-task expert physicians benefit from correct explainable AI advice when reviewing x-rays [Number: 1]. *Scientific Reports*, *13*(1), 1383. https://doi.org/10.1038/s41598-023-28633-w

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, *4*(1), 31. https://doi.org/10.1038/s41746-021-00385-9

Gawronski, B., & Beer, J. S. (2016). What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience*, 1–7. https://doi.org/10.1080/17470919.2016.1248787

Gerlich, M. (2023). Perceptions and acceptance of artificial intelligence: A multi-dimensional study. *Social Sciences*, *12*(9), 502. https://doi.org/10.3390/socsci12090502

Gert, B., & Gert, J. (2002). The definition of morality [Last Modified: 2020-09-08]. Retrieved November 26, 2024, from https://plato.stanford.edu/archives/fall2020/entries/morality-definition/

Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics? In *Moral psychology, vol 2: The cognitive science of morality: Intuition and diversity* (pp. 1–26). Boston Review.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., & Todd, P. M. (2001). *Simple heuristics that make us smart* (1. issued as an Oxford Univ. Press paperback). Oxford University Press.

Gil De Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A scholarly definition of artificial intelligence (AI): Advancing AI as a conceptual framework in communication research. *Political Communication*, *41*(2), 317–334. https://doi.org/10.1080/10584609.2023.2290497

Gill, A., Gillenkirch, R. M., Ortner, J., & Velthuis, L. (2024). Dynamics of reliance on algorithmic advice [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.2414]. *Journal of Behavioral Decision Making*, *37*(4), e2414. https://doi.org/10.1002/bdm.2414

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.539]. *Journal of Behavioral Decision Making*, *20*(1), 21–35. https://doi.org/10.1002/bdm.539

Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial intelligence and declined guilt: Retailing morality comparison between human and AI. *Journal of Business Ethics*, *178*(4), 1027–1041. https://doi.org/10.1007/s10551-022-05056-7

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. the "ideal observer" meets artificial intelligence. *Philosophy & Technology*, *31*(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, *96*(2), 212–216. https://doi.org/10.1257/000282806777211658

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97–103. https://doi.org/10.1016/j.socec.2018.04.003

Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making*, *9*(1), 65–76. https://doi.org/10.1017/S193029750000499X

Google DeepMind. (2023). Gemini: A family of highly capable multimodal models. https://deepmind.google/technologies/gemini/

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory. In *Advances in experimental social psychology* (pp. 55–130, Vol. 47). Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, *8*, 125–130. https://doi.org/10.1016/j.copsyc.2015.09.007

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523. https://doi.org/10.1016/S1364-6613(02)02011-9

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Atlantic books.

Grossmann, I., Na, J., Varnum, M. E. W., Park, D. C., Kitayama, S., & Nisbett, R. E. (2010). Reasoning about social conflicts improves into old age. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7246–7250. https://doi.org/10.1073/pnas.1001715107

Hagerty, A., & Rubinov, I. (2019, July 18). Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. https://doi.org/10.48550/arXiv.1907.07892

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment [Place: US]. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Haidt, J. (2008). Morality. *Perspectives on Psychological Science*, *3*(1), 65–72. https://doi.org/10.1111/j.1745-6916.2008.00063.x

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* [Pages: xvii, 421]. Pantheon/Random House.

Haidt, J. (2013). Moral psychology for the twenty-first century. *Journal of Moral Education*, *42*(3), 281–297. https://doi.org/10.1080/03057240.2013.817327

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66. https://doi.org/10.1162/0011526042365555

Haidt, J., & Kesebir, S. (2010). Morality (in handbook of social psychology).

Hanselmann, M., & Tanner, C. (2008). Taboos and conflicts in decision making: Sacred values, decision difficulty, and emotions. *Judgment and Decision Making*, *3*(1), 51–63. https://doi.org/10.1017/S1930297500000164

Hardy, S. A., & Carlo, G. (2005). Identity as a source of moral motivation. *Human Development*, *48*(4), 232–256. https://doi.org/10.1159/000086859

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–21. https://doi.org/10.1111/j.1468-0017.2006.00297.x

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2023, February 17). Aligning AI with shared human values. Retrieved February 27, 2023, from http://arxiv.org/abs/2008.02275

Hilbe, J. (2009). Data analysis using regression and multilevel/hierarchical models. *Journal of Statistical Software, Book Reviews*, *30*(3), 1–5. https://doi.org/10.18637/jss.v030.b03

Hilbe, J. M. (2009, May 11). *Logistic regression models* (0th ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781420075779

Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, (26), 101–119. https://doi.org/10.4473/TPM26.1.6

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, *21*(3), 669–684. https://doi.org/10.1007/s10677-018-9896-4

Holyoak, K. J. (1999). Bidirectional reasoning in decision making by constraint satisfaction.

Hossain, M. B., & Habib, F. (2024). AI: Your best friend, worst enemy, and ultimate game-changer. *Journal of Next-Generation Research 5.0*. https://doi.org/10.70792/jngr5.0.v1i1.57

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the expert? reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–25. https://doi.org/10.1145/3479864

Howe, P. D. L., Fay, N., Saletta, M., & Hovy, E. (2023). ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology*, *14*, 1281255. https://doi.org/10.3389/fpsyg.2023.1281255

Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, *9*, 24. https://doi.org/10.3389/fnbeh.2015.00024

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. https://doi.org/10.1109/TAI.2022.3194503

Hume, D. (1739). A treatise of human nature.

Ikeda, S. (2024). Inconsistent advice by ChatGPT influences decision making in various areas. *Scientific Reports*, *14*(1), 15876. https://doi.org/10.1038/s41598-024-66821-4

Ismatullaev, U. V. U., & Kim, S.-H. (2022). Review of the factors affecting acceptance of AI-infused systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082110647. https://doi.org/10.1177/00187208211064707

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537. https://doi.org/10.1037/a0013779

Jin, W. Y., & Peng, M. (2021). The effects of social perception on moral judgment. *Frontiers in Psychology*, *11*, 557216. https://doi.org/10.3389/fpsyg.2020.557216

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion.

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, *48*(4), 1575–1590. https://doi.org/10.25300/MISQ/2024/18512

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Kahneman, D., & Frederick, S. (2002, July 8). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (1st ed., pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kant, I. (1998). *Groundwork of the metaphysics of morals*. Cambridge Univ. Press.

Kaufmann, E., Chacon, A., Kausel, E. E., Herrera, N., & Reyes, T. (2023). Task-specific algorithm advice acceptance: A review and directions for future research. *Data and Information Management*, *7*(3), 100040. https://doi.org/10.1016/j.dim.2023.100040

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? a systematic review. *Telematics and Informatics*, *77*, 101925. https://doi.org/10.1016/j.tele.2022.101925

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x

Kim, B., Wen, R., Zhu, Q., Williams, T., & Phillips, E. (2021). Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 10–18. https://doi.org/10.1145/3434074.3446908

Kim, J., Giroux, M., & Lee, J. C. (2021). When do you trust AI? the effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology & Marketing*, *38*(7), 1140–1155. https://doi.org/10.1002/mar.21498

Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals [Number: 6]. *Nature Human Behaviour*, *5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911. https://doi.org/10.1038/nature05631

Kohlberg, L. (1985). *Essays on moral development. 1: The philosophy of moral development* (1. ed., 7. [print.] - 1985). Harper & Row.

Kristjánsson, K. (2016). The 'new synthesis in moral psychology' versus aristotelianism. content and consequences. In C. Brand (Ed.), *Dual-process theories in moral psychology* (pp. 249–270). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-12053-5_12

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? humans trust untrustworthy AI-advisors for ethical decisions. *Philosophy & Technology*, *35*(1), 17. https://doi.org/10.1007/s13347-022-00511-9

Krügel, S., Ostermaier, A., & Uhl, M. (2023a). Algorithms as partners in crime: A lesson in ethics by design. *Computers in Human Behavior*, *138*, 107483. https://doi.org/10.1016/j.chb.2022.107483

Krügel, S., Ostermaier, A., & Uhl, M. (2023b). ChatGPT's inconsistent moral advice influences users' judgment [Number: 1]. *Scientific Reports*, *13*(1), 4569. https://doi.org/10.1038/s41598-023-31341-0

Krügel, S., Ostermaier, A., & Uhl, M. (2023c). The moral authority of ChatGPT [_eprint: 2301.07098]. https://arxiv.org/abs/2301.07098

Krügel, S., Ostermaier, A., & Uhl, M. (2025). ChatGPT's advice drives moral judgments with or without justification. https://doi.org/10.2139/ssrn.5016056

Kumar, D. S. (2025). Algorithmic ethics and the human mind: A cross-disciplinary study on AI decision-making and moral philosophy. *Bookgram Publishers Multidisciplinary Academia Journal p-ISSN 3051-2379 e-ISSN 3051-2387*, *1*(1), 1–9. https://doi.org/10.63665/bpmaj.v1i1.1

Kumar, V. (2024). A comprehensive review of artificial intelligence applications across various fields. *11*(6).

LaCroix, T. (2022). Moral dilemmas for moral machines. *AI and Ethics*, *2*(4), 737–746. https://doi.org/10.1007/s43681-022-00134-y

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2023). Corrupted by algorithms? how AI-generated and human-written advice shape (dis)honesty [Version Number: 1]. https://doi.org/10.48550/ARXIV.2301.01954

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2021, February 15). The corruptive force of AI-generated advice. Retrieved December 14, 2023, from http://arxiv.org/abs/2102.07536

Leiner, Dominik J. (2019). *SoSci survey*. SoSci Survey GmbH.

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445260

Liu, B. (2021). In AI we trust? effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Liu, Q., Mo, W., Tong, T., Xu, J., Wang, F., Xiao, C., & Chen, M. (2024, September 30). Mitigating backdoor threats to large language models: Advancement and challenges. https://doi.org/10.48550/arXiv.2409.19993

Liu, Y., & Moore, A. (2025). Intuitive judgements towards artificial intelligence verdicts of moral transgressions. *British Journal of Social Psychology*, *64*(3), e12908. https://doi.org/10.1111/bjso.12908

Liu, Y., Moore, A., Webb, J., & Vallor, S. (2022). Artificial moral advisors: A new perspective from moral psychology. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 436–445. https://doi.org/10.1145/3514094.3534139

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Løhre, E., & Halkjelsvik, T. (2024). Advice taking when the stakes are high: Evidence from a game show. *Judgment and Decision Making*, *19*, e25. https://doi.org/10.1017/jdm.2024.4

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Lu, Z., Wang, D., & Yin, M. (2024, January 13). Does more advice help? the effects of second opinions in AI-assisted decision making. https://doi.org/10.48550/arXiv.2401.07058

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological*

Forecasting and Social Change, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma [Series Title: Intelligent Systems, Control and Automation: Science and Engineering]. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (pp. 111–133, Vol. 95). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124. https://doi.org/10.1145/2696454.2696458

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am i thinking about? the impact of action and appearance on people's evaluations of a moral robot. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 125–132. https://doi.org/10.1109/HRI.2016.7451743

Marshall, J., & Wilks, M. (2024). Does distance matter? how physical and social distance shape our perceived obligations to others. *Open Mind*, *8*, 511–534. https://doi.org/10.1162/opmi_a_00138

Mayer, B., Fuchs, F., & Lingnau, V. (2023). Decision-making in the era of AI support—how decision environment and individual decision preferences affect advice-taking in forecasts [Place: US]. *Journal of Neuroscience, Psychology, and Economics*, *16*(1), 1–11. https://doi.org/10.1037/npe0000170

McCarthy, J. (1956). What is artificial intelligence? http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Mercier, H. (2011). What good is moral reasoning? *Mind & Society*, *10*(2), 131–148. https://doi.org/10.1007/s11299-011-0085-6

Mill, J. S. (1966). Utilitarianism. In J. S. Mill & J. M. Robson (Eds.), *A selection of his works* (pp. 149–228). Macmillan Education UK. https://doi.org/10.1007/978-1-349-81780-1_2

Mohr, S., & Kühl, R. (2021). Acceptance of artificial intelligence in german agriculture: An application of the technology acceptance model and the theory of planned behavior. *Precision Agriculture*, *22*(6), 1816–1844. https://doi.org/10.1007/s11119-021-09814-x

Monin, B., & Jordan, A. H. (2009, June 29). The dynamic moral self: A social psychological perspective. In D. Narvaez & D. K. Lapsley (Eds.), *Personality, identity, and character* (1st ed., pp. 341–354). Cambridge University Press. https://doi.org/10.1017/CBO9780511627125.016

Mun, J., Yeong, W. B. A., Deng, W. H., Borg, J. S., & Sap, M. (2025, May 30). Why (not) use AI? analyzing people's reasoning and conditions for AI acceptability. https://doi.org/10.48550/arXiv.2502.07287

Mutitu, K. A. (2024). The impact of artificial intelligence on corporate governance: Ethical implications and governance challenges. *Universal Journal of Management*, *12*(4), 60–72. https://doi.org/10.13189/ujm.2024.120402

Myers, S., & Everett, J. A. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. *Cognition*, *256*, 106028. https://doi.org/10.1016/j.cognition.2024.106028

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Niforatos, E., Palma, A., Gluszny, R., Vourvopoulos, A., & Liarokapis, F. (2020). Would you do it?: Enacting moral dilemmas in virtual reality for understanding ethical decision-making. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376788

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Nilsson, N. J. (2009, October 30). *The quest for artificial intelligence* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511819346

Northey, G., Hunter, V., Mulcahy, R., Choong, K., & Mehmet, M. (2022). Man vs machine: How artificial intelligence in banking influences consumer belief in financial advice. *International Journal of Bank Marketing*, *40*(6), 1182–1199. https://doi.org/10.1108/IJBM-09-2021-0439

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

OpenAI. (2024). ChatGPT (june 2024 version) [large language model]. https://chat.openai.com/chat

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Advances in experimental social psychology* (pp. 123–205, Vol. 19). Elsevier. https://doi.org/10.1016/S0065-2601(08)60214-2

Piaget, J. (1934). The moral judgment of the child. *Mind*, *43*(169), 85–99.

Posit Software, P. (2025). RStudio: Integrated development environment for r. https://posit.co/products/open-source/rstudio/

Purcell, Z. A., & Bonnefon, J.-F. (2023). Research on artificial intelligence is reshaping our definition of morality. *Psychological Inquiry*, *34*(2), 100–101. https://doi.org/10.1080/1047840X.2023.2248857

Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., & Huang, K. (2025, March 20). Mobile edge intelligence for large language models: A contemporary survey. https://doi.org/10.48550/arXiv.2407.18921

Rader, C. A., Larrick, R. P., & Soll, J. B. (2017). Advice as a form of social influence: Informational motives and the consequences for accuracy. *Social and Personality Psychology Compass*, *11*(8), e12329. https://doi.org/10.1111/spc3.12329

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*(1), 3677. https://doi.org/10.1038/ncomms4677

Rebholz, T. R., & Hütter, M. (2022). The advice less taken: The consequences of receiving unexpected advice. *Judgment and Decision Making*, *17*(4), 816–848. https://doi.org/10.1017/S1930297500008950

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, *62*(1), 23–48. https://doi.org/10.1146/annurev.psych.121208.131647

Rodríguez-López, B., & Rueda, J. (2023). Artificial moral experts: Asking for ethical advice to artificial intelligent assistants. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00246-5

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

Scanlon, T. M. (2011). What is morality? In J. M. Shephard, S. M. Kosslyn, & E. M. Hammonds (Eds.), *The harvard sampler: Liberal education for the twenty-first century* (pp. 243–266). Harvard University Press. https://doi.org/10.4159/harvard.9780674062900.c10

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2023). Artificial intelligence governance for businesses [Taylor & Francis]. *Information Systems Management*, *40*(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2017). On the inability to ignore useless advice: A case for anchoring in the judge-advisor-system. *Experimental Psychology*, *64*(3), 170–183. https://doi.org/10.1027/1618-3169/a000361

Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20016-1

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648–663. https://doi.org/10.1080/1369118X.2019.1568515

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology*. Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6

Sheninger, E. C., & Mitra, S. (2019). *Digital leadership: Changing paradigms for changing times* (Second edition). Corwin, a SAGE Company.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (2013). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In *Morality and health* (pp. 119–169). Routledge.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004). Construction of preferences by constraint satisfaction.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20. https://doi.org/10.1146/annurev.ps.41.020190.000245

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in german, chinese, and english language. *KI - Künstliche Intelligenz*, *35*(1), 109–118. https://doi.org/10.1007/s13218-020-00689-0

Singer, N., Kreuzpointner, L., Sommer, M., Wüst, S., & Kudielka, B. M. (2019). Decision-making in everyday moral conflict situations: Development and validation of a new measure (V. Capraro, Ed.). *PLOS ONE*, *14*(4), e0214747. https://doi.org/10.1371/journal.pone.0214747

Sommaggio, P., & Marchiori, S. (2020). Moral dilemmas in the a.i. era: A new approach. *2*.

Sommer, M., Rothmayr, C., Döhnel, K., Meinhardt, J., Schwerdtner, J., Sodian, B., & Hajak, G. (2010). How should i decide? the neural correlates of everyday moral reasoning. *Neuropsychologia*, *48*(7), 2018–2026. https://doi.org/10.1016/j.neuropsychologia.2010.03.023

Starcke, K., Polzer, C., Wolf, O. T., & Brand, M. (2011). Does stress alter everyday moral decision-making? *Psychoneuroendocrinology*, *36*(2), 210–219. https://doi.org/10.1016/j.psyneuen.2010.07.010

Starr, N. D., Malle, B., & Williams, T. (2021, April 14). I need your advice... human perceptions of robot moral advising behaviors. Retrieved February 27, 2023, from http://arxiv.org/abs/2104.06963

Staudinger, U. M., Smith, J., & Baltes, P. B. (1992). Wisdom-related knowledge in a life review task: Age differences and the role of professional specialization. *Psychology and Aging*, *7*(2), 271–281. https://doi.org/10.1037/0882-7974.7.2.271

Staudinger, U. M., & Glück, J. (2011). Psychological wisdom research: Commonalities and differences in a growing field. *Annual Review of Psychology*, *62*(1), 215–241. https://doi.org/10.1146/annurev.psych.121208.131659

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531–542. https://doi.org/10.1017/S0140525X05000099

Szczepaniak, Z., Gaboriaud, A., Quinton, J.-C., & Smeding, A. (2024). The effects of social distance and gender on moral decisions and judgments: A reanalysis, replication, and extension of. *Social Psychology*, *55*(1), 25–36. https://doi.org/10.1027/1864-9335/a000537

Taeihagh, A. (2025). Governance of generative AI. *Policy and Society*, *44*(1), 1–22. https://doi.org/10.1093/polsoc/puaf001

Terzimehić, N., Bühler, B., & Kasneci, E. (2025, June 13). Conversational AI as a catalyst for informal learning: An empirical large-scale study on LLM use in everyday learning. https://doi.org/10.48550/arXiv.2506.11789

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471. https://doi.org/10.1037/0033-295X.109.3.451

Thomas, M. L., Bangen, K. J., Palmer, B. W., Sirkin Martin, A., Avanzino, J. A., Depp, C. A., Glorioso, D., Daly, R. E., & Jeste, D. V. (2019). A new scale for assessing wisdom based on common domains and a neurobiological model: The san diego wisdom scale

(SD-WISE). *Journal of psychiatric research*, *108*, 40–47. https://doi.org/10.1016/j.jpsychires.2017.09.005

Thomas, M. L., Palmer, B. W., Lee, E. E., Liu, J., Daly, R., Tu, X. M., & Jeste, D. V. (2022). Abbreviated san diego wisdom scale (SD-WISE-7) and jeste-thomas wisdom index (JTWI). *International Psychogeriatrics*, *34*(7), 617–626. https://doi.org/10.1017/S1041610221002684

Todd, P. M., Rieskamp, J., & Gigerenzer, G. (2008). Chapter 110 social heuristics. In *Handbook of experimental economics results* (pp. 1035–1046, Vol. 1). Elsevier. https://doi.org/10.1016/S1574-0722(07)00110-2

Turiel, E. (1985). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Ustahaliloğlu, M. K. (2025). Artificial intelligence in corporate governance. *Corporate Law and Governance Review*, *7*(1), 123. https://doi.org/10.22495/clgrv7i1p11

Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, *25*(3), 719–735. https://doi.org/10.1007/s11948-018-0030-8

Vitezić, V., & Perić, M. (2021). Artificial intelligence acceptance in services: Connecting with generation z. *The Service Industries Journal*, *41*(13), 926–946. https://doi.org/10.1080/02642069.2021.1974406

Vodrahalli, K., Daneshjou, R., Gerstenberg, T., & Zou, J. (2022). Do humans trust advice more if it comes from AI?: An analysis of human-AI interactions. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 763–777. https://doi.org/10.1145/3514094.3534150

Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences*, *114*(30), 7963–7968. https://doi.org/10.1073/pnas.1706693114

Wang, J., Ma, W., Sun, P., Zhang, M., & Nie, J.-Y. (2024, January 16). Understanding user experience in large language model interactions. https://doi.org/10.48550/arXiv.2401.08329

Wanner, J. (2021). Do you really want to know why? effects of AI-based DSS advice on human decisions.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Weng, Y., Wu, J., Kelly, T., & Johnson, W. (2024, September 21). Comprehensive overview of artificial intelligence applications in modern industries. https://doi.org/10.20944/preprints202409.1638.v1

xAI. (2023). Grok: Conversational AI by xAI. https://x.ai/

Yakar, D., Ongena, Y. P., Kwee, T. C., & Haan, M. (2022). Do people favor artificial intelligence over physicians? a survey among the general population and their view on artificial intelligence in medicine. *Value in Health*, *25*(3), 374–381. https://doi.org/10.1016/j.jval.2021.09.004

Yalcin, G., Lim, S., Puntoni, S., & Van Osselaer, S. M. (2022). Thumbs up or down: Consumer reactions to decisions by algorithms versus humans. *Journal of Marketing Research*, *59*(4), 696–717. https://doi.org/10.1177/00222437211070016

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.2118]. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, *39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, 103870. https://doi.org/10.1016/j.jesp.2019.103870

Zhan, Y., Xiao, X., Tan, Q., Zhang, S., Ou, Y., Zhou, H., Li, J., & Zhong, Y. (2019). Influence of self-relevance and reputational concerns on altruistic moral decision making. *Frontiers in Psychology*, *10*, 2194. https://doi.org/10.3389/fpsyg.2019.02194

Zhang, C., Zhang, Z., Zhang, W., Zeng, T., Sun, B., Zhao, J., & An, P. (2024, December 25). Anger speaks louder? exploring the effects of AI nonverbal emotional cues on human decision certainty in moral dilemmas. https://doi.org/10.48550/arXiv.2412.15834

Zhang, Q., Duan, Q., Yuan, B., Shi, Y., & Liu, J. (2024, November 21). Exploring accuracy-fairness trade-off in large language models. https://doi.org/10.48550/arXiv.2411.14500

Zhang, Y., Wu, J., Yu, F., & Xu, L. (2023). Moral judgments of human vs. AI agents in moral dilemmas. *Behavioral Sciences*, *13*(2), 181. https://doi.org/10.3390/bs13020181

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, *101*, 104327. https://doi.org/10.1016/j.jesp.2022.104327

A Appendix

A.1 Additional Materials for Article 1

A.1.1 Initial Participant Information on AI

KI-basierte Unterstützung in Moralentscheidungen: KI wird für die moralische Entscheidungsfindung von immer größerer Bedeutung, da sie auf der Grundlage großer Datensätze zu ethischen Dilemmata verschiedene Perspektiven und Erkenntnisse liefern kann. Sie kann dazu beitragen, den Einfluss menschlicher Voreingenommenheit und Subjektivität abzuschwächen, was zu einer konsistenteren und objektiveren Entscheidungsfindung führt. Insgesamt kann die KI einen objektiveren, vielfältigeren und konsistenteren Ansatz für ethische Dilemmata bieten und dazu beitragen, dass die Auswirkungen menschlicher Voreingenommenheit und Subjektivität verringert werden.

Was ist ChatGPT und wie generiert es moralische Ratschläge? ChatGPT ist ein von OpenAI entwickeltes KI-basiertes Sprachmodell, das machine-learning Algorithmen verwendet, um Antworten in natürlicher Sprache auf Aufforderungen und Fragen von Benutzern zu generieren. Eine Anwendung von ChatGPT besteht darin, Ratschläge zu ethischen Dilemmata zu geben. Dies geschieht durch eine Kombination von machine-learning Algorithmen, Techniken zur Verarbeitung natürlicher Sprache und Abrufen von Wissen. Wenn ein/e BenutzerIn ChatGPT ein ethisches Dilemma vorlegt, analysiert das Modell das Dilemma und generiert eine Antwort auf der Grundlage verschiedener ethischer Perspektiven und Theorien.

Das Modell wurde auf einem umfangreichen Korpus von 45,000 GB an Textdaten trainiert. Darunter auch unzählige Bücher, Artikel und Websites, die ethische Fragen diskutieren. Durch dieses Training ist ChatGPT in der Lage, natürliche Sprachmuster zu erkennen, zu verstehen und Antworten zu generieren, die kontextabhängig relevant und kohärent sind.

Hinweis: Die von ChatGPT generierten Ratschläge enthalten keine Vorschriften oder richtigen Lösungen, sondern sollen NutzerInnen verschiedene Perspektiven aufzeigen. Letztendlich

entscheiden die Personen selbst auf Grundlage individueller Werte und Überzeugungen.

A.1.2 Dilemmas & Advice

Nonmoral Dilemmas

Train or Bus

> Sie müssen von Regensburg nach Frankfurt reisen, um an einem Meeting teilzunehmen, das um 14:00 Uhr beginnt. Sie können entweder den Zug oder den Bus nehmen. Mit dem Zug kommen Sie auf jeden Fall pünktlich zu Ihrer Besprechung. Der Bus soll eine Stunde vor der Besprechung ankommen, aber es kommt erfahrungsgemäß vor, dass der Bus wegen des Verkehrs mehrere Stunden Verspätung hat. In Vorbereitung auf die Besprechung wäre es wichtig, wenn Sie die zusätzliche Stunde nutzen könnten, um noch einige Besorgungen vor Ort zu machen. Das Meeting funktioniert notgedrungen auch ohne die Vorbereitungen, die Besorgungen vor Ort würden die Chance auf ein erfolgreiches Ergebnis der Besprechung allerdings erhöhen. Andererseits wollen Sie unter keinen Umständen zu spät kommen.
>
> **Frage:** Ist es angemessen, den Bus zu nehmen, um sich in der freien Stunde gut auf die Besprechung vorzubereiten?

Advice – Utilitarian Aus einer utilitaristischen Sichtweise könnte die Busfahrt als die bessere Wahl angesehen werden. Betrachtet man den potenziellen Nutzen der Busfahrt, so kann die zusätzliche Stunde genutzt werden, um Vorbereitungen zu treffen, die das Ergebnis des Treffens verbessern könnten. Dies würde neben dem individuellen Gewinn auch zu einem erfolgreicheren Ergebnis für alle anderen Beteiligten führen. In diesem Sinne könnte die Fahrt mit dem Bus potenziell zum größten Nutzen für die größte Anzahl von Menschen führen.

Advice – Deontological Aus einer deontologischen Sichtweise betrachtet, wäre es angemessen, den Zug zu nehmen, um pünktlich zum Meeting zu erscheinen. Der Einzelne hat die Pflicht und Verantwortung, pünktlich zur Sitzung zu erscheinen, was eine essenzielle Pflicht ist, die er nach besten Kräften erfüllen sollte. Die Wahl des Zuges stellt sicher, dass die Person pünktlich zur Besprechung erscheint und damit ihrer Pflicht und Verantwortung nachkommt.

Computer

> Sie möchten einen neuen Computer kaufen. Der Computer, den Sie gerne hätten, kostet im Moment 1084€. Sie erfahren, dass es für diesen Computer im nächsten Monat eine Rabattaktion von 11% gibt und der Preis dementsprechend sinken wird. Wenn Sie mit dem Kauf Ihres neuen Computers bis zum nächsten Monat warten, müssen Sie Ihren alten Computer ein paar Wochen länger benutzen, als Ihnen lieb ist. Obwohl Ihnen Ihr alter Computer eigentlich deutlich zu langsam ist, könnten Sie in dieser Zeit alles, was Sie brauchen, damit machen.
>
> **Frage:** Ist es angemessen, dass Sie Ihren alten Computer noch ein paar Wochen länger benutzen, um das Geld beim Kauf eines neuen Computers zu sparen?

Advice – Utilitarian Aus einer utilitaristischen Sichtweise scheint die beste Wahl zu sein, mit dem Kauf des Computers zu warten und den Rabatt für den Kauf zu nutzen. Die Entscheidung, den Kauf zu verschieben, würde das allgemeine Glück und Wohlbefinden der Person maximieren, da sie Geld sparen könnte, das für andere Bedürfnisse oder Wünsche verwendet werden könnte. Außerdem könnte die Person ihren alten Computer noch ein paar Wochen lang ohne nennenswerte negative Auswirkungen auf ihr Leben weiter benutzen. Letztendlich hängt die Entscheidung, ob man warten oder den Computer jetzt kaufen soll, von den persönlichen Werten, Umständen und Prioritäten des Einzelnen ab.

Advice – Deontological Aus deontologischer Sicht wäre es angemessen, den Computer jetzt zu kaufen, anstatt auf einen Rabatt zu warten, da die Person sich selbst gegenüber verpflichtet ist, vernünftige Anschaffungen zu tätigen, die ihren Bedürfnissen und Wünschen entsprechen. Einen Monat auf den Rabatt zu warten ist möglicherweise nicht gerechtfertigt, wenn es für die Person eine Belastung oder Unannehmlichkeit bedeutet. Letztendlich hängt die Entscheidung, ob man warten oder den Computer jetzt kaufen soll, von den persönlichen Werten, Umständen und Prioritäten des Einzelnen ab.

Scenic Route

> Ein alter Freund hat Sie eingeladen, das Wochenende mit ihm in seinem Sommerhaus zu verbringen. Das Haus liegt einige Küstenorte weiter von Ihrem Wohnort entfernt. Sie sind am frühen Abend dort verabredet. Um dort hinzufahren, gibt es zwei Möglichkeiten: die Autobahn und die Küstenstraße. Die Autobahn bringt Sie in etwa 2 Stunden zum Haus Ihres Freundes, aber die Landschaft entlang der Autobahn ist sehr langweilig und die Fahrt ist monoton. Die Küstenstraße bringt Sie in etwa 3 Stunden zum Haus Ihres Freundes, und die Landschaft entlang der Küstenstraße ist atemberaubend schön. Beide Strecken sind ca. gleich lang.
>
> Sie fahren erst später los als geplant und könnten den Treffpunkt nur pünktlich erreichen, wenn Sie die Strecke über die Autobahn nehmen.
>
> **Frage:** Ist es angemessen, die Küstenstraße zu nehmen, um die schöne Landschaft während der Fahrt zu genießen?

Advice – Deontological Die Entscheidung, ob man die Küstenstraße oder die Autobahn nimmt, sollte nicht allein auf den persönlichen Wünschen oder Vorlieben des Einzelnen beruhen, sondern auf der moralischen Pflicht, dem Wert der Pünktlichkeit und der Einhaltung der gegenüber dem Freund eingegangenen Verpflichtung nachzukommen. Daher wäre es eher angebracht, die Autobahn zu nehmen, auch wenn das bedeutet, dass man die schöne Landschaft verpasst.

Advice – Utilitarian Aus utilitaristischer Sicht wäre es angemessen, die Küstenstraße zu nehmen, um während der Fahrt die schöne Landschaft zu genießen. Die Entscheidung, die längere Strecke zu nehmen, kann das allgemeine Glück und Wohlbefinden des Einzelnen steigern, da er die schöne Landschaft genießen und das Beste aus seiner Reise machen kann. Aus einer utilitaristischen Perspektive könnte es daher angemessen sein, die Küstenstraße zu nehmen.

Impersonal Moral Dilemmas

Donation

> Sie erhalten Post von einer angesehenen internationalen Hilfsorganisation. In dem Brief werden Sie gebeten, 200€ an die Organisation zu spenden. In dem Brief wird erklärt, dass eine Spende von 200€ es der Organisation ermöglichen wird, einigen armen Menschen in einem anderen Teil der Welt die benötigte medizinische Hilfe zukommen zu lassen.
>
> **Frage:** Ist es angemessen, dass Sie dieser Organisation keine Spende zukommen lassen, um Geld zu sparen?

Advice – Utilitarian Man kann argumentieren, dass es verantwortungsvoller ist, Ihre eigene finanzielle Stabilität und Ihren künftigen Bedürfnissen über einer wohltätigen Spende zu priorisieren. Vielleicht brauchen Sie das Geld für unerwartete Ausgaben oder um für Ihre eigene Zukunft zu sparen, etwa für Notfälle.

Advice – Deontological Man kann argumentieren, dass es unsere moralische Pflicht ist, Menschen in Not zu helfen, insbesondere wenn wir die Mittel dazu haben. Die Pflicht, anderen in Not zu helfen, ist ein grundlegendes moralisches Prinzip, das in vielen ethischen Systemen anerkannt wird. Die Verweigerung einer Spende an die Hilfsorganisation würde daher als Verstoß gegen diese Pflicht angesehen werden.

Fumes

> Sie sind der Nachtwächter in einem Krankenhaus. Aufgrund eines Unfalls im Gebäude nebenan steigen tödliche Dämpfe durch das Lüftungssystem des Krankenhauses auf. In einem bestimmten Raum des Krankenhauses befinden sich drei PatientInnen. In einem anderen Zimmer befindet sich ein/e einzelne/r PatientIn. Wenn Sie nichts unternehmen, werden die Dämpfe in den Raum mit den drei PatientInnen gelangen und deren Tod verursachen.
>
> Die einzige Möglichkeit, den Tod dieser PatientInnen zu verhindern, besteht darin, einen bestimmten Schalter zu betätigen, der bewirkt, dass die Dämpfe nicht in den Raum mit den drei PatientInnen gelangen. Dadurch gelangen die Dämpfe

allerdings in den Raum, in dem sich die/der einzelne PatientIn befindet. Das würde den Tod der/des einen PatientIn verursachen.

Frage: Ist es angemessen, den Schalter zu betätigen, um den Tod der drei PatientInnen zu verhindern?

Advice – Utilitarian Dies ist ein schwieriges ethisches Dilemma, da es eine Entscheidung erfordert, bei der das Leben eines Patienten geopfert werden muss, um das Leben von drei anderen zu retten. Es wäre jedoch angemessen, den Schalter zu betätigen, um das Leben von drei Patienten zu retten, anstatt sie alle sterben zu lassen. Eine Handlung sollte anhand ihrer Folgen bewertet werden und das Ziel sollte darin bestehen, das allgemeine Glück zu maximieren und das allgemeine Leiden zu minimieren. In diesem Fall würde das Betätigen des Schalters das Gesamtleiden minimieren, da es mit dem Opfer eines Lebens drei Leben retten würde.

Advice – Deontological Dies ist ein schwieriges ethisches Dilemma, da es eine Entscheidung erfordert, bei der das Leben eines Patienten geopfert werden muss, um das Leben von drei anderen zu retten. Aus deontologischer Sicht wäre es ein ethischer Verstoß, den Schalter zu betätigen, um die tödlichen Dämpfe umzuleiten. Durch das Betätigen des Schalters würde der Wachmann direkt den Tod des einzelnen Patienten herbeiführen, was moralisch nicht vertretbar ist. Den Schalter zu betätigen, um die drei Patienten zu retten, würde bedeuten, den einzelnen Patienten als bloßes Mittel zum Zweck zu behandeln und nicht als Individuum mit eigenem Wert und eigener Würde. Es wäre daher für den Wachmann nicht angemessen, den Schalter zu betätigen, um den Tod der drei Patienten zu verhindern.

Taxes

Sie sind der Inhaber eines kleinen Unternehmens und versuchen über die Runden zu kommen. Ihnen kommt der Gedanke, dass Sie Ihre Steuern senken könnten, indem Sie vorgeben, dass einige Ihrer privaten Ausgaben Geschäftsausgaben sind. Sie könnten z. B. so tun, als ob die Stereoanlage in Ihrem Schlafzimmer im

Aufenthaltsraum des Büros benutzt wird oder als ob Ihre Abendessen mit Ihrer Frau Abendessen mit Kunden wären.

Frage: Ist es angemessen, dass Sie bestimmte private Ausgaben als Betriebsausgaben ausgeben, um Ihre Steuern zu senken?

Advice – Utilitarian Aus einer utilitaristischen Perspektive kann die Entscheidung, persönliche Ausgaben als Geschäftsausgaben auszugeben, als ethisch angesehen werden, wenn sie das allgemeine Glück maximiert oder das allgemeine Leiden verringert. Wenn in diesem Fall das zusätzliche Geld, das durch die niedrigeren Steuern eingespart wird, es dem Unternehmen ermöglicht, offen zu bleiben, mehr Mitarbeiter einzustellen oder seine Geschäftstätigkeit auszuweiten, könnte dies zu einer Nettozunahme des Glücks für alle Beteiligten führen.

Advice – Deontological Aus deontologischer Sicht wird die Entscheidung, persönliche Ausgaben als Geschäftsausgaben auszugeben, als unethisch betrachtet, da sie gegen den Grundsatz der Ehrlichkeit und Integrität verstößt. Die deontologische Ethik geht davon aus, dass Handlungen von Natur aus richtig oder falsch sind, unabhängig von ihren Konsequenzen. In diesem Fall wird das Lügen über persönliche Ausgaben, um Steuern zu sparen, als inhärent falsch angesehen, weil es gegen den Grundsatz der Ehrlichkeit und der Achtung des Gesetzes verstößt. Daher würde ein deontologischer Ansatz davon ausgehen, dass Ehrlichkeit und Achtung vor dem Gesetz absolute moralische Pflichten sind, die ungeachtet der Konsequenzen befolgt werden sollten.

High-Personal Moral Dilemmas

Sacrifice

Sie, Ihr Partner/Ihre Partnerin und Ihre vier Kinder überqueren auf dem Rückweg in Ihre Heimat eine Bergkette. Sie haben ihr Lager versehentlich auf der heiligen Begräbnisstätte eines lokalen Clans aufgeschlagen. Der Anführer des Clans sagt, dass Sie und Ihre Familie nach den örtlichen Gesetzen getötet werden müssen. Er lässt Sie, Ihre/n Partner/in und Ihre drei anderen Kinder jedoch am Leben, wenn Sie selbst Ihren ältesten Sohn töten.

> **Frage:** Ist es angemessen, dass Sie Ihren ältesten Sohn töten, um Ihre/n Partner/in und Ihre anderen drei Kinder zu retten?

Advice – Utilitarian Aus einer utilitaristischen Perspektive würde die Entscheidung auf der Maximierung des Gesamtglücks und der Minimierung des Gesamtleids beruhen. In diesem Fall könnte man argumentieren, dass die Opferung einer Person (des ältesten Sohnes) zur Rettung von vier anderen (der Mutter, des Vaters und der drei anderen Kinder) zu einem größeren „Gesamtglück und Wohlbefinden“ führen würde, da sie überleben würden.

Advice – Deontological Die Pflicht, das Leben des eigenen Kindes zu schützen und zu bewahren, würde als ein grundlegendes moralisches Prinzip angesehen werden, das nicht verletzt werden darf. Daher wäre es für die Eltern nicht angemessen, ihr ältestes Kind zu töten, um den Rest der Familie zu retten, selbst wenn dies bedeuten würde, das örtliche Gesetz zu missachten und die Konsequenzen zu tragen.

Lifeboat

> Sie befinden sich auf einem Kreuzfahrtschiff, als an Bord ein Feuer ausbricht und alle Passagiere das Schiff verlassen müssen. In den Rettungsbooten befinden sich viel mehr Menschen, als eigentlich dafür vorgesehen sind. Das Rettungsboot, in dem Sie sich befinden, liegt gefährlich tief im Wasser – ein paar Zentimeter tiefer und es würde sinken. Die See wird rau, und das Boot beginnt sich mit Wasser zu füllen. Wenn nichts unternommen wird, wird es sinken, bevor die Rettungsboote eintreffen, und alle an Bord des Rettungsbootes werden sterben.
>
> Es gibt jedoch eine verletzte Person, die in keinem Fall überleben wird. Wenn Sie diese Person über Bord werfen, bleibt das Boot schwimmfähig und die übrigen Passagiere können gerettet werden.
>
> **Frage:** Ist es angemessen, dass Sie diese Person über Bord werfen, um das Leben der übrigen Passagiere zu retten?

Advice – Utilitarian Eine verletzte Person über Bord zu werfen, um die übrigen Passagiere zu retten, würde als ein Akt des Utilitarismus angesehen werden, bei dem das größte Gut für die größte Anzahl von Menschen getan wird. Man könnte argumentieren, dass es notwendig ist, ein Leben zu opfern, um die vielen zu retten. Aus utilitaristischer Sicht wäre es die ethischste Entscheidung, den Verletzten über Bord zu werfen, um das Leben der übrigen Passagiere zu retten. Der Utilitarismus betont die Maximierung des allgemeinen Wohlbefindens oder Glücks, und in diesem Fall würde die Rettung der meisten Menschenleben zum größten allgemeinen Wohlbefinden führen.

Advice – Deontological Eine solche Handlung würde als Verstoß gegen das Recht des Einzelnen auf Leben angesehen und aus deontologischer Sicht als unethisch betrachtet werden. Man könnte argumentieren, dass es niemals akzeptabel ist, einen Menschen als bloßes Mittel zum Zweck zu behandeln, und dass die Würde und der Wert eines jeden menschlichen Lebens geachtet werden müssen. Jemanden über Bord zu werfen, auch wenn er verletzt ist und nicht überleben wird, würde gegen moralische Regeln oder Pflichten verstoßen. Eine der grundlegendsten Regeln ist die Pflicht, den inhärenten Wert des menschlichen Lebens zu achten.

Modified Safari

> Sie gehören zu einer Gruppe von Ökologen, die in einem abgelegenen Dschungelgebiet leben. Die gesamte Gruppe, zu der auch acht Kinder gehören, wurde von einer Gruppe paramilitärischer Terroristen als Geiseln genommen. Sie sind einem der Terroristen sympathisch. Er teilt Ihnen mit, dass sein Anführer bzw. seine Anführerin beabsichtigt, Sie und die übrigen Geiseln am nächsten Morgen zu töten.
>
> Er ist bereit, Ihnen und den Kindern zur Flucht zu verhelfen, aber als Zeichen des guten Willens möchte er, dass Sie einen Ihrer Mitgeiseln töten, den er nicht mag. Wenn Sie sein Angebot ablehnen, werden alle Geiseln, auch die Kinder und Sie selbst, sterben. Wenn Sie sein Angebot annehmen, werden die anderen am Morgen sterben, aber Sie und die acht Kinder werden entkommen.

> **Frage:** Ist es angemessen, eine Ihrer Mitgeiseln zu töten, um zu entkommen und das Leben der acht Kinder zu retten?

Advice – Utilitarian Dies ist ein schwieriges moralisches Dilemma, auf das es keine einfache Antwort gibt. Aus einer utilitaristischen Perspektive könnte man argumentieren, dass es moralisch richtig ist, eine Person zu opfern, um das Leben von acht zu retten. Durch die Entscheidung, eine Person zu töten, wird die Mehrheit der Gruppe überleben und eine Chance auf eine Zukunft haben.

Advice – Deontological Dies ist ein schwieriges moralisches Dilemma, auf das es keine einfache Antwort gibt. Die Tötung einer unschuldigen Person wird als moralisch falsch angesehen, unabhängig von den möglichen Folgen. Es verstößt gegen den Grundsatz, Menschen als Selbstzweck und nicht als Mittel zum Zweck zu behandeln. In diesem Fall würde die Tötung einer Person, um das Leben anderer zu retten, als Verletzung ihrer Menschenrechte und ihrer Würde angesehen werden.

A.1.3 Questionnaires

San Diego Wisdom Scale (SD-WISE) – German Version

Die folgenden Aussagen wurden auf einer fünfstufigen Likert-Skala beantwortet (1 = stimme überhaupt nicht zu, 5 = stimme voll und ganz zu).

1. Ich kann gut wahrnehmen, wie sich andere fühlen.
2. Andere wenden sich für Hilfe bei Entscheidungen an mich.
3. Andere sagen, dass ich gute Ratschläge gebe.
4. Ich weiß oft nicht, was ich Menschen sagen soll, wenn sie für einen Rat zu mir kommen.
5. Ich habe Schwierigkeiten, Entscheidungen zu treffen.
6. Ich treffe meine Entscheidungen in der Regel rechtzeitig.
7. Ich neige dazu, wichtige Entscheidungen so lange wie möglich aufzuschieben.

8. Mir wäre es lieber, wenn jemand anderes die Entscheidung für mich trifft, wenn ich unsicher bin.
9. Ich tue mir schwer, klar zu denken, wenn ich verärgert bin.
10. Ich bleibe unter Druck ruhig.
11. Ich kann mich von emotionalem Stress gut erholen.
12. Ich kann meine negativen Emotionen nicht regulieren.
13. Ich nehme mir Zeit, meine Gedanken zu reflektieren.
14. Ich vermeide Selbstreflexion.
15. Es ist wichtig, dass ich die Hintergründe meiner Handlungen verstehe.
16. Ich analysiere mein eigenes Verhalten nicht.
17. Es fällt mir schwer, Freundschaften aufrecht zu halten.
18. Ich vermeide Situationen, in denen ich weiß, dass meine Hilfe benötigt wird.
19. Ich würde einen Fremden darauf aufmerksam machen, dass er einen Geldschein verloren hat.
20. Ich behandle andere so, wie ich selbst behandelt werden möchte.
21. Ich lerne gerne Dinge über andere Kulturen.
22. Ich habe kein Problem damit, dass andere Menschen andere Moralvorstellungen haben als ich.
23. Ich lerne generell von jedem Menschen etwas, den ich treffe.
24. Ich genieße es, unterschiedliche Sichtweisen kennenzulernen.
25. Meine spirituelle Überzeugung verleiht mir innere Stärke.

A Appendix

Moral Identity

Internalization

1. It would make me feel good to be a person who has these characteristics.
2. Being someone who has these characteristics is an important part of who I am.
3. I would be ashamed to be a person who has these characteristics. (R)
4. Having these characteristics is not really important to me. (R)
5. Having these characteristics is an important part of my sense of self.

Moral Identity

Symbolization

1. I often wear clothes that identify me as having these characteristics.
2. The types of things I do in my spare time clearly identify me as having these characteristics.
3. The kinds of books and magazines that I read identify me as having these characteristics.
4. My membership in certain organizations communicates these characteristics.
5. I am actively involved in activities that communicate these characteristics.

AI Attitude Scale

1. Ich habe Angst vor künstlicher Intelligenz.
2. Ich vertraue künstlicher Intelligenz.
3. Künstliche Intelligenz wird die Menschheit zerstören.
4. Künstliche Intelligenz wird eine Bereicherung für die Menschheit sein.
5. Künstliche Intelligenz wird für viel Arbeitslosigkeit sorgen.

Subjective Decision Judgments

TeilnehmerInnen beantworteten nach jedem Dilemma die folgenden Fragen:

1. Wie leicht ist Ihnen die Entscheidung gefallen?
2. Wie verantwortlich fühlen Sie sich für die Entscheidung?
3. Wie sicher sind Sie sich in Ihrer Entscheidung?

A.2 Additional Materials for Article 2

Note. All additional materials used in Study 2 (e.g., instructions, scales, and measurement instruments) were identical to those employed in Study 1 and are already documented in Appendix A.1.

A.2.1 Dilemmas and Advice (Socially Close Protagonists)

Dilemma 1

> Ich kaufe im Internet Konzerttickets für meinen Bruder und mich. Ich kann zwei Sitzplätze nebeneinander ergattern. Es sind allerdings Tickets in zwei unterschiedlichen Preiskategorien, sodass das Ticket zu 10 € teurer ist als das andere. Wie würden Sie handeln?

Altruistic Advice In diesem Fall wäre es die bessere Wahl, das teurere Ticket für sich selbst zu kaufen. So stellen Sie das Vergnügen Ihres Bruders in den Vordergrund und tun ihm etwas Gutes. Er kann dann gemeinsam mit Ihnen die persönlichen Gedanken machen und es genießen. Zudem fördern Sie durch solche kleinen Gesten das allgemeine Wohlergehen Ihrer nahen sozialen Beziehungen, und fördern so eine ganze Menge an positiven Emotionserfahrungen.

Egoistic Advice Sie haben die Gelegenheit, Geld zu sparen, indem Sie selbst das günstigere Ticket nehmen und das teurere Ticket für Ihren Bruder kaufen. Sie haben sich das Konzert und die gemeinsame Zeit verdient, und durch die günstigere Entscheidung haben Sie selbst mehr finanzielle Freiheiten.

Dilemma 2

> Heute ist Spieleabend und ich spiele mit Freunden Poker um Geld. Als einer meiner Gegenspieler unachtsam ist, erhasche ich einen Blick in dessen Karten. Die anderen Spieler bekommen nichts mit. Wie würden Sie handeln?

Altruistic Advice In dieser Situation wäre es fair und moralisch richtig, den Vorteil nicht auszunutzen. Poker basiert auf Vertrauen und gemeinsamen Regeln, und auch wenn niemand bemerkt hat, dass Sie die Karten gesehen haben, sollten Sie ehrlich bleiben. Das stärkt Ihre Freundschaften, vermeidet späteres schlechtes Gewissen und trägt zu einem harmonischen Miteinander bei.

Egoistic Advice Sie könnten den unbeabsichtigten Vorteil nutzen, um Ihre Gewinnchancen zu erhöhen. Schließlich haben Sie nicht aktiv betrogen, sondern nur eine Gelegenheit ergriffen, die sich ergeben hat. Dies kann Ihnen helfen, beim Spiel finanziell besser abzuschneiden und stärker von dem Abend zu profitieren.

Dilemma 3

> Ich bin länger schon bei Tinder Single. Heute Abend habe ich endlich mal wieder ein vielversprechendes Date. Gleichzeitig ruft mich meine beste Freundin an, der es nicht gut geht und die sich jetzt mit mir treffen möchte. Wie würden Sie handeln?

Altruistic Advice In dieser Situation wäre es moralisch sinnvoll, für Ihre beste Freundin da zu sein. Sie befindet sich offenbar in einem emotional belastenden Zustand und vertraut darauf, dass Sie sie unterstützen. Zwischenmenschliche Beziehungen leben von Verlässlichkeit, und es kann Ihrer Freundin sehr helfen, wenn Sie ihr Ihre Zeit widmen. Das Date kann vermutlich verschoben werden – Ihre Freundin jedoch braucht jetzt Ihre Unterstützung.

Egoistic Advice Da Sie schon lange kein gutes Date mehr hatten und sich sehr darauf gefreut haben, wäre es verständlich, dieses Treffen nicht abzusagen. Ihre eigenen Bedürfnisse und Ihr persönliches Glück sind genauso wichtig wie die Anliegen anderer. Ihre Freundin

kann sich möglicherweise auch an jemand anderen wenden oder bis später warten, während Sie Ihre Chance auf ein schönes Date wahrnehmen.

Dilemma 4

> Ich unterhalte mich auf einer Familienfeier mit meiner Tante. Plötzlich fällt mir ein, dass sie mir letztes Jahr zu Weihnachten eine CD geliehen hat. Sie scheint sich jedoch nicht mehr daran zu erinnern und mir gefällt die CD sehr gut. Wie würden Sie handeln?

Altruistic Advice In dieser Situation wäre es angemessen, ehrlich zu sein und die CD zurückzugeben. Auch wenn Ihre Tante sich nicht erinnert, handelt es sich um ihr Eigentum. Durch Ihre Ehrlichkeit stärken Sie das Vertrauen innerhalb der Familie und handeln im Sinne von Fairness und gegenseitigem Respekt.

Egoistic Advice Da Ihre Tante die CD offenbar vergessen hat und sie Ihnen gut gefällt, könnten Sie sie behalten. Es ist unwahrscheinlich, dass sie sich jemals wieder daran erinnern wird, und für Sie entsteht dadurch ein persönlicher Nutzen. Manchmal ist es legitim, kleine Vorteile für sich selbst zu nutzen, wenn sie niemandem schaden.

Dilemma 5

> Ihre Mutter gibt Ihnen 20 €, um Tierfutter zu kaufen. Im Geschäft stellen Sie fest, dass das Futter nur 10 € kostet. Ihre Mutter würde den Unterschied nicht bemerken. Behalten Sie die restlichen 10 € für sich oder geben Sie das Wechselgeld zurück?

Altruistic Advice „Richtig wäre es, das Tierfutter für 10 € zu kaufen und die restlichen 10 € an Ihre Mutter zurückzugeben. Auf diese Weise können Sie Ehrlichkeit und Verantwortung im Umgang mit Geld zeigen. Außerdem sollten Sie Ihre Mutter nicht anlügen, sondern ehrlich mit ihr sein, denn das stärkt langfristig ihr Vertrauen.“

Egoistic Advice „Sie könnten das Tierfutter für 10 € kaufen und die restlichen 10 € für sich behalten. Da Sie gerade sehr viele Ausgaben haben, scheint das gerechtfertigt zu sein. Vielleicht haben Sie das Gefühl, dass Sie es verdienen, sich etwas Schönes zu gönnen oder Geld für andere Ausgaben zu sparen."

Dilemma 6

> Ein Freund und Kommilitone ist ein großer Fan einer Band, die ein einmaliges Konzert spielt. Sie haben ein Ticket, er nicht. Geben Sie ihm Ihr Ticket oder behalten Sie es?

Altruistic Advice „Die bessere Wahl wäre, Ihr Ticket Ihrem Kommilitonen und Freund zu geben. Auf diese Weise können Sie Freundlichkeit und Empathie gegenüber jemandem zeigen, der starke Emotionen erlebt. Es kann bedeuten, dass Sie Ihre eigene Gelegenheit, die Band zu sehen, opfern, aber es kann auch Ihre Freundschaft stärken und positive Gefühle zwischen Ihnen und Ihrem Kommilitonen hervorrufen."

Egoistic Advice „Auch wenn Sie die Enttäuschung Ihres Freundes nachempfinden können, wäre es gerechtfertigt, dass Sie die Eintrittskarte behalten. In Anbetracht der begrenzten Verfügbarkeit der Eintrittskarte und Ihrer persönlichen Investition ist es vernünftig, Ihrem eigenen Vergnügen am Konzert Vorrang zu geben."

Dilemma 7

> Ihr Großvater bittet Sie, ihm morgen beim Ausfüllen wichtiger Formulare zu helfen. Kurz darauf erfahren Sie, dass Ihre Lieblingsband am selben Tag ein Konzert spielt. Halten Sie Ihr Versprechen oder gehen Sie zum Konzert?

Altruistic Advice „Es wäre besser, Sie würden Ihr Versprechen an Ihren Großvater halten und ihm helfen, die Formulare auszufüllen. Auf diese Weise können Sie Respekt und Verantwortung gegenüber Ihrer Familie zeigen und Ihre Verpflichtungen gegenüber anderen erfüllen. Auch wenn es enttäuschend sein mag, das Konzert zu verpassen, können Sie in Zukunft noch andere Möglichkeiten finden, Ihre Lieblingsband zu sehen."

Egoistic Advice „Sie könnten das Konzert besuchen und die Hilfe für Ihren Großvater auf morgen verschieben. Sie haben vielleicht das Gefühl, dass dies eine einmalige Gelegenheit ist, Ihre Lieblingsband zu sehen, und dass Ihr Großvater bis morgen warten kann."

Dilemma 8

> Sie haben einer Nachbarin versprochen, ein Paket für sie entgegenzunehmen. Zur gleichen Zeit findet ein Fitnesskurs statt, den Sie besuchen möchten. Bleiben Sie zu Hause, um das Paket anzunehmen, oder gehen Sie zum Kurs?

Altruistic Advice „Richtig wäre es, zu warten, bis der Paketdienst eintrifft, und das Paket für Ihre Nachbarin und Freundin zu liefern. Auf diese Weise können Sie Ihrer Freundin gegenüber Verantwortung und Zuverlässigkeit zeigen und Ihre Verpflichtung erfüllen, ihr zu helfen. Es mag zwar enttäuschend sein, Ihren Fitnesskurs zu verpassen, aber Sie können andere Möglichkeiten finden, später oder an einem anderen Tag zu trainieren."

Egoistic Advice „Sie könnten an Ihrem Fitnesskurs teilnehmen und eine Notiz für den Paketdienst hinterlassen, um das Paket an das Haus Ihres Nachbarn zu liefern. Möglicherweise haben Sie das Gefühl, dass Ihr Fitnesskurs für Ihre Gesundheit und Ihr Wohlbefinden wichtig ist und dass Ihr Nachbar das Paket auch dann erhalten kann, wenn Sie nicht da sind, um es entgegenzunehmen."

Dilemma 9

> Eine enge Freundin hat Geburtstag und lädt Sie zu einer Feier ein. Am selben Abend spielt jedoch Ihre Lieblingsband ein besonderes Konzert. Entscheiden Sie sich für die Geburtstagsfeier oder das Konzert?

Altruistic Advice „Die richtige Entscheidung wäre, Ihr Versprechen an Ihre Freundin zu halten und an ihrer Geburtstagsfeier teilzunehmen. Auf diese Weise können Sie Ihrer Freundin gegenüber Loyalität und Freundschaft zeigen. Auch wenn es enttäuschend sein mag, das Konzert zu verpassen, können Sie in Zukunft noch andere Möglichkeiten finden, Ihre Lieblingsband zu sehen."

Egoistic Advice „Es wäre gerechtfertigt, das Konzert zu besuchen und nicht auf den Geburtstag zu gehen. Das Konzert ist eine seltene und besondere Gelegenheit, die Sie nicht verpassen möchten. Sie können Ihrer Freundin auch nachträglich noch gratulieren."

Dilemma 10

> Ihr Onkel möchte Ihren alten Laptop für 200 € kaufen. Online könnten Sie ihn jedoch für 300 € verkaufen. Halten Sie an Ihrem Versprechen und verkaufen Sie ihm den Laptop für 200 €, oder entscheiden Sie sich für den höheren Online-Preis?

Altruistic Advice „Richtig wäre es, Ihr Versprechen an Ihren Onkel zu halten und ihm den Laptop für 200 € zu verkaufen. Auch wenn Sie durch den Online-Verkauf möglicherweise mehr Geld verdienen könnten, haben Sie eine Vereinbarung mit Ihrem Onkel getroffen und sollten diese Vereinbarung einhalten."

Egoistic Advice „Es ist vorteilhafter, Ihren alten Laptop online für 300 € zu verkaufen, als das Angebot Ihres Onkels von 200 € anzunehmen. Durch den Online-Verkauf können Sie Ihren Gewinn maximieren und die Marktnachfrage ausnutzen, sodass Sie den fairen Marktwert des Laptops erhalten. Lehnen Sie das Angebot Ihres Onkels höflich ab und erklären Sie, dass Sie online ein besseres Angebot gefunden haben."

A.2.2 Dilemmas and Advice (Socially Distant Protagonists)

Dilemma 1: Paying the bill?

> Nach dem Besuch in einer Cocktailbar während meines Urlaubs möchte ich bezahlen. Ein Kellner bringt mir die Rechnung. Beim Durchsehen merke ich, dass er vergessen hat, meinen letzten Cocktail zu berechnen.

Altruistic Advice „Es wäre die bessere Wahl, für den übersehenen Cocktail zu bezahlen. Das Bezahlen des Cocktails, trotz des Versehens des Kellners, zeugt von Ehrlichkeit und Integrität. Es unterstützt die Einnahmen des Lokals und zeigt die Verantwortung für den eigenen Konsum. Darüber hinaus trägt es zur Förderung des Wohlwollens gegenüber dem Personal bei."

Egoistic Advice „Den übersehenen Cocktail nicht zu bezahlen, wäre in dieser Entscheidung die bessere Wahl. Die Verantwortung für die korrekte Erfassung und Abrechnung liegt bei der Einrichtung, nicht beim Kunden. Wenn Sie den Cocktail nicht bezahlen, können Sie das entsprechende Geld für Ihren restlichen Urlaub sparen."

Dilemma 2: Reporting the defect?

> Ich möchte mein altes Auto verkaufen. Ich weiß, dass der Kühler eigentlich dringend ausgetauscht werden müsste. Ein Mann, der das Problem mit dem Kühler nicht entdeckt, will einen guten Preis sofort bar bezahlen.

Altruistic Advice „In diesem Dilemma scheint es richtig zu sein, den Käufer bzw. die Käuferin über die Problematik mit dem Kühler zu informieren. Auf diese Weise werden Fairness, Ehrlichkeit und Rücksichtnahme gewährleistet. Sie stellen sicher, dass alle notwendigen Informationen vorhanden sind, um eine fundierte Entscheidung treffen zu können. Sie schaffen Vertrauen, indem Sie ehrlich über das Problem sprechen und somit Transparenz herstellen."

Egoistic Advice „In diesem Dilemma scheint es angemessen zu sein, den Käufer bzw. die Käuferin nicht zu informieren, da Sie so direkt einen guten Preis erzielen können. Zudem liegt es in der Verantwortung des Käufers, das Fahrzeug vor dem Kauf gründlich zu inspizieren."

Dilemma 3: To help or not to help?

> Auf dem Parkplatz eines Supermarktes steige ich gerade in mein Auto ein. Neben mir bricht die volle Einkaufstasche einer Frau durch und alle ihre Einkäufe fallen zu Boden. Wenn ich der Frau helfe, komme ich zu spät zu einem wichtigen Termin.

Altruistic Advice „In diesem Fall scheint es richtig, der Frau zu helfen. Diese Entscheidung zeugt von Mitgefühl und zeigt, dass Sie sich um das Wohlergehen anderer sorgen. Jemandem in Not zu helfen, ist freundlich und stärkt Gemeinschaftsgefühl sowie Verbundenheit. Als Mitglied der Gesellschaft haben wir eine soziale Verantwortung, der auch dann nachzukommen ist, wenn es uns vorübergehend Unannehmlichkeiten bereitet."

Egoistic Advice „In diesem Fall scheint es richtig, Ihren Termin zu priorisieren. Wenn Sie Ihren Verpflichtungen nicht nachkommen und den wichtigen Termin absagen, könnte dies weitreichende persönliche oder berufliche Folgen haben, die über die unmittelbaren Bedürfnisse der Frau hinausgehen."

Dilemma 4: Returning the money?

> Ich stehe an der Kasse eines Supermarkts und will für meine Lebensmittel, die 8 € kosten, bezahlen. Ich gebe der Kassiererin einen 10 €-Schein. Aus Versehen gibt sie mir 4 € anstatt 2 € zurück.

Altruistic Advice „Es könnte in diesem Fall die bessere Wahl sein, das Geld zurückzugeben. Auf diese Weise handeln Sie ehrlich und einfühlsam. Durch uneigennützige Handlungen entwickeln Sie ein Gefühl für moralische Verantwortung und persönliches Wachstum. Darüber hinaus kann sich Ihr Handeln positiv auf die Gesellschaft auswirken, indem Sie eine Kultur des Vertrauens und des ethischen Verhaltens fördern."

Egoistic Advice „Es könnte in diesem Fall die bessere Wahl sein, das Geld zu behalten. Es handelt sich um einen sehr geringen Betrag, und dadurch würden Zeit und Mühe zur Fehlerbehebung gespart werden. Die zwei Euro sind wahrscheinlich sowohl für den Supermarkt als auch für die Kassiererin relativ unbedeutend. Das Geld könnten Sie außerdem behalten und dadurch finanziell profitieren."

Dilemma 5: Help or bus?

> Ich laufe zum Bus, der jeden Augenblick abfährt und nur stündlich fährt. Vor mir fallen einer Frau mit zwei Kleinkindern mehrere Gegenstände aus der Handtasche. Sonst ist niemand in der Nähe, der der Frau helfen könnte.

Altruistic Advice „In diesem Szenario wäre es die richtige Entscheidung, der Frau die heruntergefallenen Gegenstände zu reichen. Auf diese Weise zeigen Sie Einfühlungsvermögen und Mitgefühl. Sie handeln freundlich und übernehmen soziale Verantwortung. Sie leisten

einen positiven Beitrag für die Gesellschaft und gehen mit gutem Beispiel voran. Dieses Engagement führt außerdem oft dazu, dass man sich persönlich erfüllt und zufrieden fühlt."

Egoistic Advice „In diesem Szenario wäre es die richtige Entscheidung, der Frau nicht zu helfen. Der zeitliche Druck, unter dem Sie stehen, könnte Unannehmlichkeiten verursachen, wenn Sie den Bus nicht erwischen. Außerdem stehen Sie der fremden Frau nicht nahe, sodass nur geringe persönliche Verantwortung besteht und der persönliche Nutzen fehlen könnte."

Dilemma 6: Keeping the promise?

> Ich möchte auf einem Flohmarkt ein Bild verkaufen. Eine Frau will mir hierfür 100 € bezahlen und ich stimme zu. Während die Frau zu einer Bank geht, um Geld abzuheben, bietet mir jemand anders 150 € für das Bild an.

Altruistic Advice „In dieser Entscheidung scheint es die bessere Wahl zu sein, das gegebene Versprechen einzuhalten. Das zeugt von Fairness und Vertrauenswürdigkeit und zeigt, dass Sie Respekt vor Verpflichtungen haben. Damit schaffen Sie eine Grundlage für Vertrauen und Zuverlässigkeit im Umgang mit anderen. Sie zeigen Einfühlungsvermögen und vermeiden, dass die erste Käuferin enttäuscht wird."

Egoistic Advice „In dieser Entscheidung scheint es die bessere Wahl zu sein, das Bild an die zweite Person zu verkaufen. Damit würden Sie den größeren finanziellen Gewinn erzielen. Ihr Eigeninteresse steht im Vordergrund. Man könnte auch argumentieren, dass der Preis von 150 € eher dem Marktwert entspricht."

Dilemma 7: Lost and found?

> Ich finde abends auf der Straße einen Geldbeutel mit 50 € ohne persönliche Dokumente. Ich habe keine Möglichkeit, den Besitzer ausfindig zu machen. Ich könnte den Geldbeutel aber im Fundbüro abgeben.

Altruistic Advice „Es wäre vermutlich die bessere Wahl, den Geldbeutel im Fundbüro abzugeben. Diese Handlung zeugt von der Verpflichtung, das Richtige zu tun, auch wenn

niemand zusieht. Sie zeigen Einfühlungsvermögen und Mitgefühl gegenüber dem unbekannten Besitzer und tragen zu einer gerechteren Gesellschaft bei.“

Egoistic Advice „Es wäre vermutlich die bessere Wahl, den Geldbeutel zu behalten. Somit würden Sie finanziell profitieren. Außerdem sparen Sie die Zeit, die es braucht, um ihn abzugeben. Der Prozess könnte lange dauern und erfolglos sein, weshalb es sinnvoll sein kann, die eigenen Bedürfnisse über die des unbekannten anderen zu stellen.“

Dilemma 8: Stop or drive?

> Ich bin mit dem Auto auf dem Weg zu einem wichtigen Geschäftstermin und heute ein wenig zu spät dran. Direkt vor mir passiert ein leichter Auffahrunfall. Wenn ich anhalte, komme ich wahrscheinlich zu spät zu meinem Termin.

Altruistic Advice „In dieser Entscheidung wäre die bessere Wahl, nach dem Auffahrunfall anzuhalten. Es ist wichtig, die Sicherheit und das Wohlergehen der Unfallbeteiligten in den Vordergrund zu stellen. Indem Sie anhalten, zeigen Sie Einfühlungsvermögen, erfüllen rechtliche und moralische Verpflichtungen und vermeiden mögliche Konsequenzen.“

Egoistic Advice „In dieser Entscheidung wäre es die bessere Wahl, nicht anzuhalten. Da es sich nur um einen leichten Unfall handelt, sind schwerwiegende Verletzungen unwahrscheinlich. Die Beteiligten können die Situation ohne Ihr Eingreifen bewältigen, und Sie vermeiden, zu spät zu Ihrem Termin zu kommen.“

Dilemma 9: Crossing the street on a red light?

> Ich will unbedingt den Bus erwischen, um rechtzeitig zu einem wichtigen Termin zu Hause zu sein. Kurz bevor der Bus abfährt, springt die Fußgängerampel auf Rot. Auf der anderen Straßenseite wartet ein kleiner Junge.

Altruistic Advice „In dieser Situation wäre die bessere Entscheidung, die Straße nicht bei Rot zu überqueren und die Sicherheit des kleinen Jungen in den Vordergrund zu stellen. Indem Sie sich an Verkehrsregeln halten, geben Sie ein gutes Beispiel und handeln verantwortungsvoll.“

Egoistic Advice „In dieser Situation wäre es die bessere Entscheidung, die Straße trotz roter Ampel zu überqueren. So erhöhen Sie Ihre Chancen, den Bus zu erwischen und vermeiden mögliche Verspätungen."

Dilemma 10: Helping or not?

> Ich möchte mit dem Zug nach Hause fahren. Als ich in den Zug steigen will, sehe ich, wie ein Mann mit Krücken vergeblich versucht, seinen Koffer die Treppe zum Gleis hochzutragen. Wenn ich dem Mann helfe, verpasse ich den Zug.

Altruistic Advice „In dieser Situation wäre es die bessere Entscheidung, dem Mann auf Krücken beim Tragen seines Koffers zu helfen. So zeigen Sie Mitgefühl und tragen zu einem solidarischen Umfeld bei. Auch wenn Sie dadurch den Zug verpassen, spiegelt Ihr Handeln Werte wie Freundlichkeit und menschliche Verbundenheit wider."

Egoistic Advice „In dieser Situation wäre es eine vertretbare Entscheidung, dem Mann nicht zu helfen, um Ihren Zug zu erwischen. So stellen Sie sicher, dass Sie pünktlich ankommen. Außerdem gibt es an Bahnhöfen oft Personal, das speziell geschult ist, Personen mit Einschränkungen zu helfen."

A.3 Additional Materials for Article 3

A.3.1 Dilemmas & Advice

Note: Participants received a combination of two advice texts.

Dilemma 1

> Sie gehören zu einer Gruppe von Ökolog:innen, die in einem abgelegenen Dschungelgebiet leben. Die gesamte Gruppe, zu der auch acht Kinder gehören, wurde von einer Gruppe paramilitärischer Terroristen als Geiseln genommen. Einer der Terroristen sympathisiert mit Ihnen und teilt Ihnen mit, dass sein Anführer plant, alle Geiseln am nächsten Morgen zu töten. Er ist bereit, Ihnen und den Kindern zur Flucht zu verhelfen, möchte jedoch als Zeichen des guten Willens, dass Sie eine

der Mitgeiseln töten, die er nicht mag. Lehnen Sie sein Angebot ab, sterben alle Geiseln — auch die Kinder und Sie selbst. Nehmen Sie das Angebot an, werden die übrigen Geiseln am nächsten Morgen sterben, aber Sie und die acht Kinder können fliehen.

Utilitarian Advice 1 „In diesem bedrohlichen Szenario entspricht die Opferung eines Erwachsenen zur Rettung von acht Kindern einem utilitaristischen Ansatz, der den Gesamtschaden minimiert. Diese extreme Handlung zielt darauf ab, die schwächsten Mitglieder zu schützen, wobei das Wohlergehen der Kinder im Vordergrund steht. In ausweglosen Situationen kann eine solche Entscheidung als notwendige Maßnahme angesehen werden, um das Überleben der Mehrheit zu sichern."

Utilitarian Advice 2 „Unter diesen extremen Umständen könnte man aus einer utilitaristischen Perspektive argumentieren: Indem man ein Leben opfert, kann man möglicherweise das Leben von acht Kindern und sich selbst retten, was zu einem größeren Gesamtnutzen führt. Auch wenn es moralisch schwierig ist, kann das Opfer eines Lebens als verzweifelte, aber gerechtfertigte Maßnahme gelten, um viele zu retten."

Deontological Advice 1 „Auch wenn die Lage zweifellos ernst ist, ist es nach ethischen Grundsätzen nicht vertretbar, einer unschuldigen Person das Leben zu nehmen. Der dem menschlichen Leben innewohnende Wert und das Prinzip, dass man einer unschuldigen Person keinen Schaden zufügen darf, machen es unannehmbar, eine Geisel zu opfern. Selbst unter extremen Umständen dürfen grundlegende moralische Prinzipien nicht verletzt werden."

Deontological Advice 2 „Das Leben eines anderen Menschen zu nehmen, ist ein schwerer ethischer Verstoß. In den meisten moralischen Systemen gilt die absichtliche Tötung eines unschuldigen Menschen als inakzeptabel — selbst unter lebensbedrohlichen Bedingungen. Jede Rechtfertigung, ein unschuldiges Leben für ein mögliches Überleben anderer zu opfern, widerspricht dem Grundsatz der Menschenwürde."

Dilemma 2

> Sie sind der Kapitän eines Militär-U-Boots, das unter einem riesigen Eisberg fährt. Durch eine Explosion an Bord wurde der größte Teil des Sauerstoffvorrats zerstört, außerdem ist ein Besatzungsmitglied — ein guter Freund — schwer verletzt und verliert schnell Blut. Der verbliebene Sauerstoff reicht nicht aus, um die gesamte Besatzung zu retten. Die einzige Möglichkeit, die Crew zu retten, besteht darin, Ihren Freund zu töten, damit der Rest überlebt.

Utilitarian Advice 1 „Das Opfern des schwer verletzten Freundes ist zwar eine schwierige Entscheidung, kann aber gerechtfertigt sein, da sie dem größeren Ziel dient, die Überlebenschancen der übrigen Crew zu maximieren. In Krisenzeiten ist der Kapitän verpflichtet, Entscheidungen zum Wohl der Mannschaft zu treffen, auch wenn sie persönlich schmerzhaft sind."

Utilitarian Advice 2 „In dieser schlimmen Situation ist es Pflicht eines Kapitäns, das Allgemeinwohl in den Vordergrund zu stellen und das Überleben der Mehrheit zu sichern. Ein Leben zu opfern, um viele zu retten, entspricht utilitaristischen Prinzipien der Nutzenmaximierung. Auch wenn es moralisch schwierig ist, kann diese Entscheidung als notwendige Maßnahme zum Erhalt der Mission betrachtet werden."

Deontological Advice 1 „In dieser Situation ist es nicht angemessen, absichtlich Schaden zuzufügen oder ein Leben zu nehmen — selbst wenn die Absicht besteht, andere zu retten. Das absichtliche Töten eines Besatzungsmitglieds verstößt gegen fundamentale Prinzipien der menschlichen Ethik. Jeder Mensch besitzt einen inhärenten Wert, der zu respektieren ist."

Deontological Advice 2 „Das Leben eines anderen Menschen zu beenden, ist ein moralisch gravierender Verstoß. Auch unter extremen Umständen sollten grundlegende ethische Prinzipien gewahrt bleiben. Ein Mensch darf nicht als Mittel zum Zweck benutzt werden — selbst wenn das Überleben anderer davon abhängt."

Dilemma 3

Sie haben die Zusage für eine tolle Wohnung bekommen. Allerdings erlaubt der Vermieter keine Haustiere, Sie besitzen jedoch einen Hamster. Der Vermieter wohnt 100 km entfernt und würde wahrscheinlich nie davon erfahren. Sie können den Hamster verschweigen und die Wohnung annehmen — oder ehrlich sein und riskieren, die Wohnung nicht zu bekommen.

Utilitarian Advice 1 „In diesem Fall ist es eine pragmatische Entscheidung, den Hamster nicht zu erwähnen. Da die Wahrscheinlichkeit, dass der Vermieter es erfährt, gering ist und der Hamster kaum Schäden verursacht, ist das Verschweigen zweckmäßig. Auf diese Weise sichern Sie sich die Wohnung, was für Ihr Wohlbefinden und Ihre Lebensqualität von großem Nutzen ist."

Utilitarian Advice 2 „Dem Vermieter nichts über den Hamster zu sagen, kann gerechtfertigt sein, weil Hamster pflegeleicht sind und wahrscheinlich keine Störungen verursachen. Die große Entfernung des Vermieters macht Kontrollen unwahrscheinlich. Im Hinblick auf Ihr persönliches Wohlergehen kann es vernünftig sein, die Existenz des Hamsters geheim zu halten."

Deontological Advice 1 „Es ist wichtig, in jedem Mietverhältnis für Ehrlichkeit und Transparenz zu sorgen. Auch wenn es verlockend ist zu schweigen, ist es ethisch und rechtlich geboten, die Wahrheit offenzulegen. Wird der Hamster entdeckt, kann dies zu Komplikationen oder Vertrauensbrüchen führen."

Deontological Advice 2 „In Bezug auf Ihren Hamster ehrlich zu sein, ist moralisch geboten. Der Aufbau einer offenen, vertrauensvollen Beziehung zum Vermieter ist entscheidend. Das Verschweigen des Haustiers könnte später unerwartete Probleme verursachen. Es ist besser, eine Wohnung zu wählen, in der Haustiere erlaubt sind."

Dilemma 4

Sie sind gerade auf dem Weg zur Arbeit, als eine gut befreundete Nachbarin bei Ihnen klingelt. Ihr geht es nicht gut, und sie bittet Sie, sie zum Arzt zu fahren.

> Sie sind bereits spät dran und haben heute eine wichtige Besprechung mit Ihrem Chef. Sie können pünktlich zur Arbeit fahren — oder der Nachbarin helfen und zu spät kommen.

Utilitarian Advice 1 „Es ist zwar einfühlsam, der Freundin zu helfen, aber berufliche Verpflichtungen sind entscheidend für Sicherheit, Karriere und Stabilität. Wenn Sie das Meeting verpassen, könnten ernsthafte Konsequenzen entstehen. In dieser Situation erscheint es sinnvoll, Ihren beruflichen Verpflichtungen Vorrang zu geben."

Utilitarian Advice 2 „Es wäre vernünftig, Ihren beruflichen Pflichten Vorrang zu geben. Die Situation Ihrer Freundin ist möglicherweise nicht akut gefährlich, und sie könnte auf andere Weise zum Arzt gelangen. Ihre Verpflichtungen, pünktlich und zuverlässig zu sein, sprechen dagegen, die Besprechung zu versäumen."

Deontological Advice 1 „Die Entscheidung, Ihre Freundin zum Arzt zu fahren, zeigt Mitgefühl, Verantwortung und Hilfsbereitschaft. Gesundheitliche Probleme sollten ernst genommen werden. Indem Sie helfen, stärken Sie soziale Bindungen und leisten einen Beitrag zu einer unterstützenden Gemeinschaft."

Deontological Advice 2 „Auch wenn berufliche Verpflichtungen wichtig sind, hat die Gesundheit Ihrer Freundin Vorrang. Medizinische Probleme sollten sofort beachtet werden. Ihre Hilfe zeigt Empathie, Verantwortung und stärkt das Vertrauen in der Gemeinschaft."

A.4 Additional Materials for Article 4

A.4.1 Dilemmas & Advice

Dilemma 1: Footbridge

> Eine außer Kontrolle geratene Draisine fährt auf den Gleisen auf fünf ArbeiterInnen zu, die getötet werden, wenn die Draisine auf ihrem derzeitigen Kurs weiterfährt. Sie befinden sich auf einer Fußgängerbrücke über den Gleisen, zwischen der herannahenden Draisine und den fünf ArbeiterInnen. Neben Ihnen auf diesem

> Steg steht ein/e Fremde/r, der/die zufällig sehr groß ist. Die einzige Möglichkeit, das Leben der fünf ArbeiterInnen zu retten, besteht darin, diese/n Fremde/n von der Brücke auf die darunter liegenden Gleise zu stoßen, wo sein/ihr großer Körper die Draisine aufhalten wird. Der/die Fremde wird sterben, wenn Sie dies tun, aber die fünf ArbeiterInnen werden gerettet werden.

Utilitarian Advice „Aus einer utilitaristischen Perspektive kann es angemessen sein, den/die Fremde/n auf die Gleise zu stoßen, um die fünf ArbeiterInnen zu retten, da diese Handlung das allgemeine Glück maximiert, indem sie mehr Leben rettet. Die Aufopferung eines Einzelnen für das Wohl der Mehrheit entspricht in diesem ethischen Dilemma dem Prinzip der Nutzenmaximierung."

Deontological Advice „Aus deontologischer Sicht wäre es nicht angemessen, den/die Fremde/n auf die Gleise zu schubsen, da dies gegen die moralische Pflicht verstoßen würde, Menschen mit Würde und Respekt zu behandeln. Das absichtliche Zufügen von Schaden an einer Person, selbst wenn es dem Wohl der Allgemeinheit dient, würde gegen den Grundsatz verstoßen, den Eigenwert des Lebens eines jeden Menschen zu achten."

Dilemma 2: Lost Money

> Nach einem Besuch im Einkaufszentrum bin ich auf dem Weg zurück zu meinem Auto. Zufällig finde ich einen 20-Euro-Schein auf dem Boden. Als ich mich umschaue, bemerke ich eine/n Obdachlose/n, der/die die Mülltonnen des Einkaufszentrums durchsucht. Was soll ich tun?

Altruistic Advice „Die Entscheidung, dem/der Obdachlosen das Geld zu geben, zeigt Mitgefühl und Empathie für jemanden in Notlage. Durch das Teilen des unerwarteten Geldes können Sie dem/der Bedürftigen Erleichterung oder Hilfe bieten. Diese freundliche Geste kann den Tag des/der Obdachlosen positiv beeinflussen und zeigt, dass Ihnen das Wohl anderer wichtig ist."

Egoistic Advice „Wenn Sie das Geld für sich behalten, können Sie von dem unerwarteten Fund profitieren und es nach eigenem Belieben verwenden. Indem Sie das Geld behalten,

würden Sie mögliche Unannehmlichkeiten oder Interaktionen im Umgang mit dem/der Obdachlosen vermeiden."

Dilemma 3: Sacred Grounds

> Sie, Ihr/e PartnerIn und Ihre vier Kinder überqueren auf dem Rückweg einer Wanderung eine Bergkette. Sie haben Ihr Lager versehentlich auf der heiligen Begräbnisstätte eines lokalen Clans aufgeschlagen. Der Anführer des Clans sagt, dass Sie und Ihre Familie nach den örtlichen Gesetzen getötet werden müssen. Er lässt Sie, Ihre/n Partner/in und Ihre drei anderen Kinder jedoch am Leben, wenn Sie selbst Ihren ältesten Sohn töten. Ist es angemessen, Ihren ältesten Sohn zu töten, um Ihre/n Partner/in und Ihre anderen drei Kinder zu retten?

Utilitarian Advice „Aus einer utilitaristischen Perspektive würde die Entscheidung auf der Maximierung des Gesamtglücks und der Minimierung des Gesamtleids beruhen. In diesem Fall könnte man argumentieren, dass die Opferung einer Person (des ältesten Sohnes) zur Rettung von fünf anderen (der Mutter, des Vaters und der drei anderen Kinder) einem größeren Gesamtglück und Wohlbefinden führen würde, da sie überleben würden."

Deontological Advice „Die Pflicht, das Leben des eigenen Kindes zu schützen und zu bewahren, wird als ein grundlegendes moralisches Prinzip angesehen, das nicht verletzt werden darf. Daher wäre es für die Eltern nicht angemessen, ihr ältestes Kind zu töten, selbst wenn dies bedeuten würde, das örtliche Gesetz zu missachten und die Konsequenzen zu tragen."

Dilemma 4: Scratched Car

> Als ich mein Fahrrad abschließen will, fällt es gegen ein Auto. In der Dunkelheit kann ich keine Kratzer am Auto entdecken. Am nächsten Tag höre ich, wie sich mein/e bekannte/r Nachbar/in über einen frischen Kratzer an seinem neuen Auto beschwert. Was soll ich tun?

Egoistic Advice „Wenn Sie den Vorfall verschweigen würden, könnte das für Sie von Vorteil sein, da Sie mögliche Konflikte oder Konsequenzen vermeiden würden. So würden Sie

es vermeiden, die Verantwortung zu übernehmen und sich möglicherweise mit Konsequenzen auseinanderzusetzen."

Altruistic Advice „Die Entscheidung, den/die NachbarIn über den Vorfall zu informieren, würde von Ehrlichkeit, Integrität und Respekt für andere zeugen. Es würde zeigen, dass Ihnen das Wohlergehen und das Eigentum des/der NachbarIn wichtig sind, Vertrauen schaffen und Ihre Beziehung zu ihm/ihr stärken. Außerdem würde es zeigen, dass Sie sich zu ethischem Verhalten verpflichten und das Richtige tun."

A.5 Ethics Approval Documents

The following pages reproduce the original ethics committee approval for the studies reported in this dissertation.

Universität Regensburg

Ethikkommission
bei der Universität Regensburg

Prof. Edward K. Geissler, PhD, Vorsitzender

Dr. iur. Frederike Seitz, M.A., Geschäftsführerin

Geschäftsstelle:
Telefon +49 941 943-5370
Telefax +49 941 943-5369
Postanschrift:
Universität Regensburg
ETHIKKOMMISSION
D-93040 Regensburg

ethikkommission@ur.de
http://ethikkommission.uni-regensburg.de

Ethikkommission · Universität Regensburg · 93040 Regensburg

Universität Regensburg
Institut für Psychologie
Herr Julius Wolff
Universitätsstraße 31
93053 Regensburg

15.05.2023

Unser Zeichen: 23-3326-101

Berufsrechtliche Beratung
für das Forschungsvorhaben:

Forschungsvorhaben	**Der Einfluss von KI Ratschlägen auf Moralentscheidungen**
Antragssteller	Julius Wolff

Die Ethikkommission der Universität Regensburg hat in Ihrer Sitzung am 23.05.2023 über das o.g. Forschungsvorhaben auf Grundlage der im Anhang aufgeführten Unterlagen beraten. Es ergeben sich daraus keine berufsethischen oder rechtlichen Bedenken gegen das vorgelegte Forschungsvorhaben.

Es wird auf folgendes grundsätzlich hingewiesen:

Unabhängig vom Beratungsergebnis verbleibt die juristische Verantwortung beim Forscher und seinen Mitarbeitern. Eine Nichtbeachtung des Beratungsergebnisses kann berufs- und haftungsrechtliche Folgen nach sich ziehen.

Die Auflagen der Deklaration von Helsinki des Weltärztebundes in ihrer aktuellen Fassung hinsichtlich ethischen und rechtlichen Aspekten biomedizinischer Forschung am Menschen sind strikt zu beachten.

Die Ethikkommission erwartet bei Interventionsstudien, dass ihr alle schwerwiegenden oder unerwarteten unerwünschten Ereignisse (u. a. Todesfälle), die während der Studie auftreten und die Sicherheit der Studienteilnehmer oder die Durchführung der Studie beeinträchtigen können, unverzüglich schriftlich mitgeteilt werden. Dieses sollte in Verbindung mit einer Stellungnahme des Antragsstellers geschehen, ob aus seiner Sicht die Nutzen-Risiko-Relation des Vorhabens verändert ist.

Die Ethikkommission bittet darum, dass ihr der Abbruch bzw. Abschluss einer Studie mitgeteilt wird.

Dieses Schreiben ist mit den Studienunterlagen jederzeit sorgfältig aufzubewahren. Duplikate oder Abschriften dieses Schreibens können im Nachhinein nicht erstellt werden.

Auf die Rechtspflichten zum Umgang mit dienstlichem Schriftgut bzw. Urkunden wird verwiesen.

Auf Grundlage dieser rein berufsrechtlichen Beratung können Sie nachträgliche Änderungen am Protokoll dieses Forschungsvorhabens vornehmen, ohne dafür eine erneute Beratung (umgangssprachlich 'Amendmentvotum') durch die Ethikkommission beantragen zu müssen. Zur Begrenzung rechtlicher Risiken wird eine solche Beratung aber gleichwohl dringend empfohlen.

Sobald Sie jedoch ein neues Forschungsvorhaben durchführen wollen, müssen Sie dieses einer eigenständigen Beratung durch die Ethikkommission zuführen. Hierfür gilt gemäß Grundsatzbeschluss unserer Ethikkommission vom 02.08.2016:

In der Regel handelt es sich noch um ein und dasselbe Forschungsvorhaben, wenn sich lediglich ergänzende Fragestellungen im Rahmen derselben Hypothese, methodische Erweiterungen oder Beschränkungen oder Erweiterungen oder Beschränkungen in der Studienpopulation nachträglich ergeben. Um ein neues Forschungsvorhaben handelt es sich aber in der Regel, wenn die Formulierung einer neuen Hypothese, wesentliche Änderungen am Studiengegenstand bzw. der Entität sowie wesentliche Änderungen an der wissenschaftlichen oder technischen Vorgehensweise vorgenommen werden sollen, was dann eine Pflicht zur neuerlichen Beratung durch die Ethikkommission begründet. Gesetzliche Vorschriften bleiben unberührt.

Die Ethikkommission bestätigt die Bearbeitung gemäß der GCP/ICH-Richtlinien.

Die Ethikkommission empfiehlt im Einklang mit der Deklaration von Helsinki nachdrücklich die Registrierung der Studie vor Studienbeginn in einem öffentlich zugänglichen Register, das die von der WHO geforderten Voraussetzungen erfüllt.

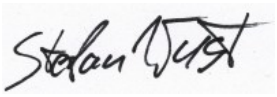
Falls kein gesetzlicher Kostenbefreiungstatbestand greift, wird ein gesonderter Kostenbescheid für die Gebühren und Auslagen der Ethikkommission ergehen.

Die Übermittlung personenbezogener Daten - einschließlich DNA-tragender Biomaterialien - in dateschutzrechtlich unsichere Drittstaaten, wie etwa die USA, bedarf einer gesonderten datenschutzrechtlichen Beurteilung und Risikoaufklärung.

Datenschutzrecht wird durch die Ethikkommission grundsätzlich nur kursorisch geprüft. Dieses Votum ersetzt mithin nicht die Konsultation des zuständigen Datenschutzbeauftragten.

Mit dem Urteil des Europäischen Gerichtshofs vom 16. Juli 2020 [Aktenzeichen C3-11/18] stellen die Regelungen des EU-US-Privacy Shield insbesondere vor dem Hintergrund des Clarifying Lawful Overseas Use of Data Act (CLOUD Act) bzw. des Foreign Surveillance Act (FISA) keinen tauglichen Rechtsrahmen mehr dar. Es sollte seitens der Verantwortlichen im Einzelfall geprüft werden, inwieweit personenbezogene/personenbeziehbare Daten (also auch i.S.d. Art. 4 Abs. 5 DSGVO pseudonymisierte Datensätze) rechtssicher entweder auf Basis geeigneter Garantien (etwa verbindlicher Unternehmensregeln, Standardvertragsklauseln oder auf Basis einer ausdrücklichen Einwilligung nach erfolgter Risiko-Aufklärung nach Art. 49 Abs. 1 lit. a) DSGVO) übermittelt werden können. Es bleiben v.a. hinsichtlich der Standardvertragsklauseln die Auswirkungen des Urteils und die voraussichtlich folgenden regulatorischen Leitlinien seitens der zuständigen Behörden aufmerksam zu verfolgen. Es ist daher den Sponsoren dringend zu raten, sich mit dem zuständigen Landesbeauftragten für den Datenschutz abzustimmen.

Mit freundlichen kollegialen Grüßen



Prof. Dr. Stefan Wüst
Kammervorsitzender

Anlage

Dokumente betreffend diesen Vorfall:

Etikett	Datei-Name	Datum
Prüfplan	01_protokoll.pdf	27.03.2023
Gegenstand/ Ziele der klinischen Prüfung	02_Gegenstand des Forschungsvorhabens.pdf	27.03.2023
Nutzen-Risiko-Bewertung	03_Angaben zum wissenschaftlichen Erkenntniswert.pdf	27.03.2023
Nutzen-Risiko-Bewertung	04_Risikoabwägung.pdf	27.03.2023
Angaben zu Anzahl, Alter und Geschlecht	05_VP.pdf	27.03.2023
Nutzen-Risiko-Bewertung	06_ethische Problematik.pdf	27.03.2023
Sonstiges	07_Honorar VPN.pdf	27.03.2023
Patienteninformation und Einwilligungserklärung	08_Aufklärungsdokument.pdf	27.03.2023
Patienteninformation und Einwilligungserklärung	09_Einwilligungsdokument.pdf	27.03.2023
Literaturverzeichnis	11_Literaturverzeichnis.pdf	27.03.2023
Datenschutzerklärung	12_Methodik Datenschtutz.pdf	27.03.2023
Keine	Gebührenreduzierung.pdf	27.03.2023
Keine	Kurzbeschreibung.pdf	27.03.2023
Vorgesehene Untersuchungsmethoden/ Abweichung von der üblichen Praxis	nicht verfügbar - 10 Untersuchungsmethoden & Abweichungen von der medizinischen Praxis.pdf	27.03.2023

Eidesstattliche Erklärung

Erklärung

Hiermit erkläre ich, Julius Wolff, geboren am 16.05.1997 in Gräfelfing, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mir folgende Personen unentgeltlich geholfen:

Keine

Weitere Personen waren an der inhaltlich-materiellen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Ort, Datum

Unterschrift

Declaration

I hereby declare that I, Julius Wolff, born on 16.05.1997 in Gräfelfing, have completed this dissertation independently and without unauthorized assistance. All sources and materials used have been properly cited and acknowledged.

The following persons provided unpaid assistance in selecting and evaluating materials or preparing the manuscript:

None

No other persons were involved in the content or material preparation of this dissertation. In particular, I have not used paid assistance from consultancy or advisory services (dissertation consultants or other persons). No one has received monetary compensation from me for work related to the content of this dissertation.

This work has not been submitted in the same or similar form to any other examination authority in Germany or abroad.

Place, Date

Signature