Journal of Artificial Intelligence Research 82 (2025) 1581-1628 Submitted 07/2024; published 03/2025

Computational Machine Ethics: A Survey

Tammy Zhong TAMMY.ZHONG@UNSW.EDU.AU
Yang Song YANG.SONG1@UNSW.EDU.AU
Raynaldio Limarga R.LIMARGA@STUDENT.UNSW.EDU.AU
Maurice Pagnucco MORRI@CSE.UNSW.EDU.AU
School of Computer Science and Engineering
The University of New South Wales, Sydney, NSW 2052, AUSTRALIA.

Abstract

Computational Machine Ethics (CME) is an interdisciplinary field that integrates moral philosophy into an agent's decision-making process, contributing to the broader domain of Artificial Intelligence Ethics. Technological advancements have transformed the world, where technology has become an integral part of society, progressively given more autonomy in making judgments within various domains in our lives. Inevitably, issues of ethics come into play in these judgments, making ethical decision-making in machines an increasingly critical problem to solve. This survey provides an overview of CME, highlighting the breadth of directions and the use of techniques within the field. We also provide some background on the ethical dimension before introducing our taxonomy used to categorise and detail the variety of existing approaches from a more technical perspective. Finally, we identify limitations in the research and suggest potential open challenges for future work.

1. Introduction

From news about Hanson Robotics' robot Sophia saying that "(She) will destroy humans"[1], a Chess robot breaking a child's finger[2] to discussions about the morals of OpenAI's ChatGPT-3 very soon after its public release[3], there has been and continue to be many more stories and discussions on the internet and in the media about the potential negative impact of Artificial Intelligence (AI). Regardless of whether stories portrayed in the media are true or not, they bring fear and negatively impact public trust. These may hinder the future of AI research and development, limiting the potential AI technologies could have in our society (Anderson & Anderson, 2007). *AI Ethics* has thus become an important field of research made up of two interrelated areas; AI, and Ethics. It is about ensuring the responsible design, development and validation of AI technologies and systems as well as the use of AI itself in addressing ethical concerns (Zhu, Xu, Lu, Governatori, & Whittle, 2022).

Computational Machine Ethics (CME) is a subfield of AI Ethics where the focus is on implementing and ensuring the moral behaviours of cognitive machines given some ethical theories and rules for the machines to follow. Although a relatively new research area that has evolved mainly in the last 15 years with the advances in AI technology (McLaren, 2006;

1. https://www.thefridaytimes.com/2022/09/06/facing-fear-the-coming-of-the-terminator/
2. https://www.theguardian.com/sport/2022/jul/24/chess-robot-grabs-and-breaks-finger-of-seven-year-old-opponent-moscow
3. http://www.mindfulmarketing.org/1/post/2023/01/an-ethics-professor-tests-the-morals-of-chatgpt.html

Edmond et al., 2022), it is an important field that unveils and examines inevitable problems for further advancements and adoption of AI in our society in the future. Through research on incorporating ethical components in AI and enabling ethics to be computational, we may be able to maintain public trust and maximise the extent and benefits which AI may bring to our society. Additionally, research in CME may also make contributions to the field of ethics (Anderson & Anderson, 2007).

Current approaches in CME are very diverse in terms of their approaches and span into various directions with differing goals including *identifying* what is ethical and unethical, *deciding* the most ethical choice within a situation, and *advising* or *mimicking* human's ethical decision making. Some contributions are on a conceptual level whilst others propose more methodological solutions such as formalisations and simulations. The aim of this survey paper is to provide an overview of approaches in CME and present our personal stance on the field.

A broadly-scoped survey presented by Ji et al. (2024) provides an overview of research in *AI Alignment* where systems aim to behave in line with human intentions and values. Although the survey overlaps with a subset of works in CME, it has less of a focus on ethics. There have also been surveys that specifically examine the entirety of CME. Recent work by Otterbacher and Manolopoulos (2023) discusses the bigger picture, discussing the evolution path of CME whilst Yu et al. (2018); Winfield, Michael, Pitt, and Evers (2019); Cervantes et al. (2020); Tolmeijer, Kneer, Sarasua, Christen, and Bernstein (2020); Nallur (2020), review the different CME approaches with more technical detail. Some others adopt an even more specific focus within CME. For instance, Talat et al. (2022) delves into the technical domain of machine learning of ethical judgments, Torras (2024) provides an overview of ethics specifically in social robots and Rossi and Mattei (2019) studies two specific examples that serve as representative instances of general directions within the field. Although not a review on the state of the art in CME from a technical perspective, Poszler, Portmann, and Lütge (2024) surveys experts from AI, cognitive sciences and philosophy to provide alternative insights on CME research.

Our work is inspired by existing surveys on CME (Yu et al., 2018; Winfield et al., 2019; Cervantes et al., 2020; Tolmeijer et al., 2020), but it differs in the way existing methods are presented and categorised. We collate and summarise five of the ongoing and significant questions in the field and, based on these questions, we introduce a new taxonomy to systematically examine these studies at a high level. Given the broad nature of this field, this paper is not intended to be exhaustive but rather to provide a representative overview of the field and the various directions explored.

We note that the main focus of this survey will be on methods developed to incorporate ethics for a single agent. We recognise that there are more studies on ethical decision-making in multi-agent systems such as (Rodriguez-Soto, Rodriguez-Aguilar, & Lopez-Sanchez, 2022). We briefly touch on a few as we attempt to map out the breadth of the field. We also want to acknowledge the existence of works in relation to the analysis and evaluation of human ethical decision-making and choices in ethical dilemmas, many through social media texts (Wilson, 2019; Awad et al., 2022; Liu, 2022; Nguyen et al., 2022; Surendran et al., 2022; Jentzsch, Schramowski, Rothkopf, & Kersting, 2019). These are not the focus of this paper and so will not be discussed further. Furthermore, although relevant to creating socially responsible AI, we will not be looking at issues such as algorithmic fairness and bias. More

information about such considerations can be found in the work by Cheng, Mosallanezhad, Sheth, and Liu (2021). The following sections will organise the literature into a taxonomy and the survey will conclude with a discussion of our views on the area with some open challenges and potential future directions.

2. Taxonomy

Recall that CME integrates ethical considerations into the decision-making processes of machines. In this section, we propose the **Source-Decision-Evaluation taxonomy** (Figure 1), designed to simulate the sequential nature of decision-making in the ethical context. As the name suggests, we will examine the ***source*** of information used to guide the ethical decision-making process, such as the adopted ethical policies and datasets used. Utilising this information as a basis or input for a system, different approaches then employ distinct methodologies to arrive at an ethical ***decision***. Lastly, within the field, certain studies ***evaluate*** other approaches or their own methods using diverse qualitative and quantitative metrics. In addition, a comprehensive breakdown of the subcategories and examined works in this survey is provided in Figure 2. Please note that the categorisation of literature in this paper, along with its subcategories introduced later, may not be strictly mutually exclusive. Certain works may fall under multiple categories, however, we have chosen to discuss them within specific section(s) that we believe best represents their primary contribution(s). It is important to acknowledge that categorising approaches involves some subjectivity; distinctions such as "bottom-up" versus "hybrid" may vary depending on individual perspectives and interpretation.

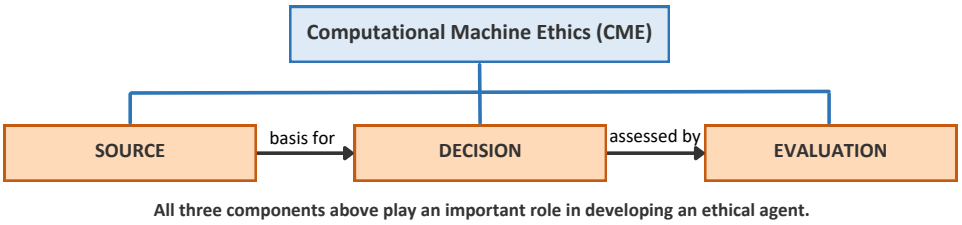

Figure 1: Overview of Source-Decision-Evaluation taxonomy.

Our taxonomy is inspired by the following five (5) significant and ongoing questions frequently encountered in AI and Ethics, and in particular within CME discussions. These questions are identified through our analysis of the literature as crucial focal points within the field. The taxonomy we propose offers a lens through which we can examine the works in the field of CME in relation to these key questions, providing a structured framework for understanding and addressing them:

1. How much autonomy can and should we enable for AI agents to make ethical judgements? (Dyrkolbotn, Pedersen, & Slavkovik, 2017; Moor, 2006).
2. What does it mean to be "ethical"? (Driver, 2006) **(SOURCE).**
3. Who should be the decision-maker(s) of what this definition should be? (Discussed in many works such as (L. A. Dennis, Fisher, Slavkovik, & Webster, 2016)) **(SOURCE).**

4. Given some definition, how can one translate such ethics and implement it in an AI agent? (Anderson & Anderson, 2007) **(DECISION).**

5. Assuming some ethical decision-making abilities, how do we evaluate/measure and ensure that these agents are consistently performing in accordance to the expected level of ethics. (Langman, Capicotto, Maddahi, & Zareinia, 2021) **(EVALUATION).**

Although beyond the scope of this survey, we would like to acknowledge the significance of the first question for further contemplation. Notably, Moor's (Moor, 2006) four levels of ethical agents holds relevance in this context:

- **Ethical-impact agents** do not have mechanisms within them to enable any kind of ethical considerations. They are simply technology which leads to an ethical impact. The example given by Moor is how robotic camel jockeys replaced young children freeing them from slavery in Qatar.

- **Implicit ethical agents** are agents who totally avoid unethical outcomes as the internal functions of the agent does not involve any unethical behaviour or it is avoided by programmer's hard-coded instructions. They have no ability to think "ethically". An example of such agent is an automated teller machine where money transactions should be honest. However, the need to be honest is not encoded in the machine but rather it has simply been programmed to perform these transactions accurately.

- **Explicit ethical agents** are those agents who have the ability to represent ethical theories and principles and perform analysis based on given information. A similar non-ethical example is a Chess game where a representation of the board is provided and the agent is able to determine the legal moves and compute its next appropriate move.

- **Full ethical agents** are similar to humans in that we are capable of making ethical judgments and justifying them. Additionally, we possess qualities such as free will, consciousness, and intentionality.

At the current stage of research in CME, as technological advances are not yet at the stage where machines can execute human-like thinking processes and behaviours to a substantial extent, we are mainly interested in building *explicit ethical agents* which is also the focus of this survey.

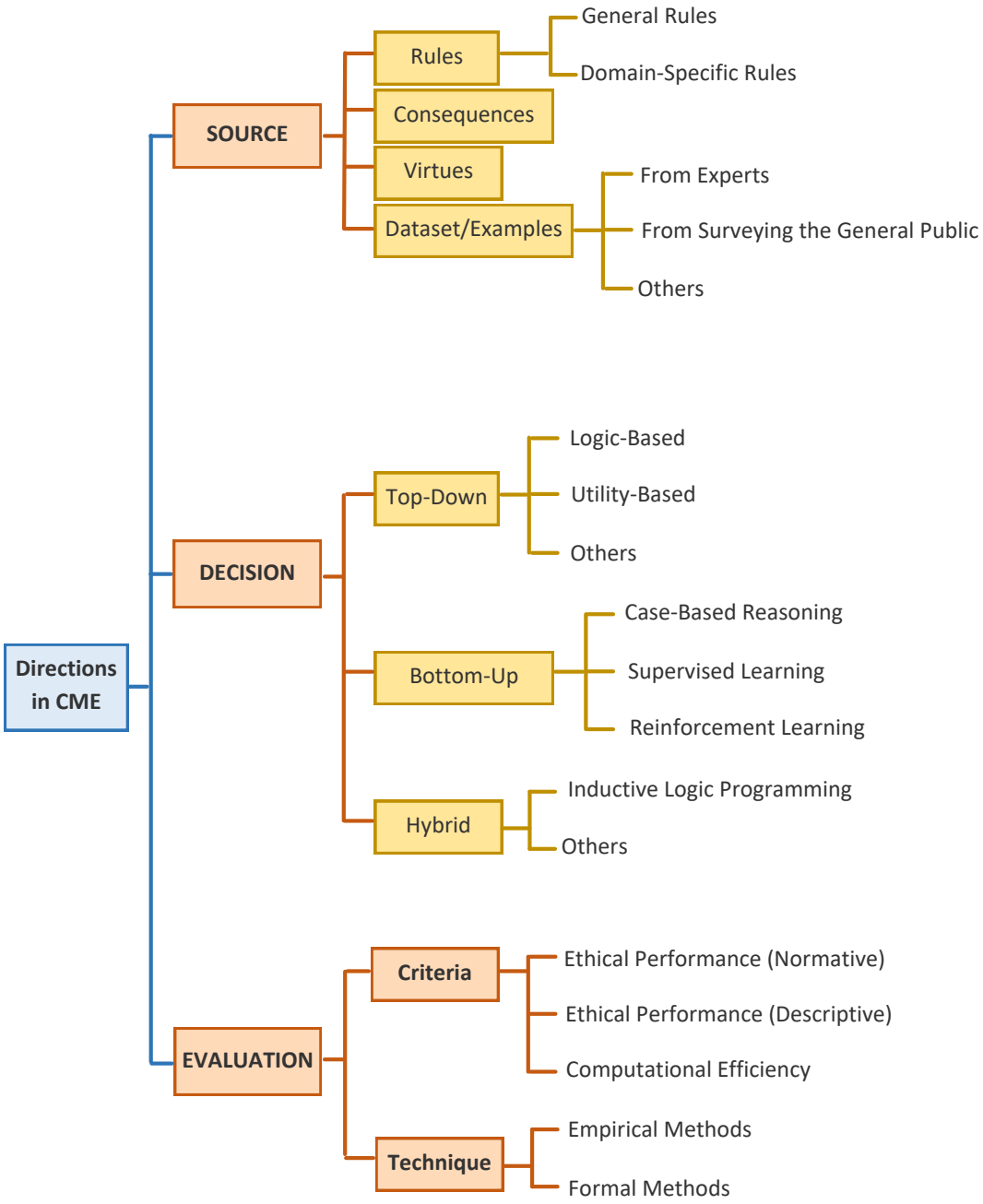

Figure 2: Overview of the Source-Decision-Evaluation taxonomy.

3. SOURCE for Guidance in Ethical Decision-Making—The Ethical Foundations

In human decision-making, especially in ethical contexts, various factors come into play, such as certain rules or guidelines, including ethical ones, our existing knowledge, past

experiences, acquired skills, and the influence of others. Similarly, when considering ethical decision-making in machines, a fundamental question arises: what should influence their ethical decisions? Therefore, it is crucial to examine what researchers have adopted to guide ethical decision-making in machines within CME. However, before doing so, we will first examine the fundamental ethical theories that serve as the foundation for decision criteria and processes, which will be discussed later in Sections 4 and 5.

Ethics involves the examination of behaviours deemed "good" or "bad", and "right" or "wrong", both for individuals and society. The main areas of ethics often discussed are *normative ethics* (also known as prescriptive ethics), which focuses on the creation of theories and criteria for governing ethical behaviour; *applied ethics*, which studies how we should act in specific domains of life; and *metaethics*, which goes beyond topics in normative ethics and applied ethics, and explores the origins and foundations of ethical principles (Dimmock & Fisher, 2017). There also exists the idea of *descriptive ethics*, often viewed as a contrast to normative ethics, which examines and describes existing practices of groups, societies, or cultures (Gert & Gert, 2020). The four distinct branches of ethics are depicted in Figure 3. However, for the purpose of this survey, our focus is on normative ethics and descriptive ethics, as they align more closely with the works in CME. Metaethics and applied ethics, while recognised as important branches, are beyond the scope of this survey. The following subsection details the major subcategories in normative ethics. Descriptive ethics primarily focuses on how people actually behave ethically and make ethical judgments rather than determining what is objectively right or wrong.

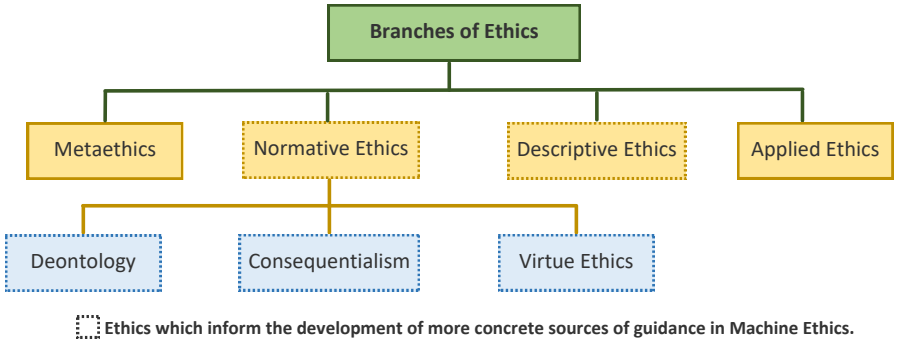

Figure 3: Branches of ethics.

3.1 Ethical Theories within Normative Ethics

As illustrated in Figure 3, normative ethics encompasses three primary ethical theories: consequentialism, deontology, and virtue ethics. However, we would also like to recognise that an alternate perspective exists among ethicists who embrace a *particularist* view. According to this viewpoint, all ethical theories are regarded as practical rule-of-thumb guidelines, rejecting the idea of universal principles or theories that can be universally applied. Instead, particularists argue for a case-by-case evaluation of each ethical situation and which approach is appropriate for that situation.

Consequentialism Consequentialism involves evaluating one's actions based on the consequences they bring about. The central idea is to maximise overall well-being. *Utilitarianism*

is a well-known example of consequentialism (Sinnott-Armstrong, 2022). According to utilitarianism, an action is considered ethically right if it contributes to overall well-being or maximises some form of utility. When faced with a range of actions to choose from, this theory asserts that the optimal choice is the one that leads to the greatest net overall well-being for all.

Deontology In deontology, an action is considered morally good if it adheres to specific moral rules or duties that are relevant to the given context. Deontological ethics can be classified into two types: agent-centred and patient-centred (Alexander & Moore, 2021). Agent-centered deontological theories involve obligations that are directed towards a specific agent, meaning that these obligations may not apply to other individuals. For instance, an individual may have specific obligations towards their own child, but these obligations would not necessarily extend to another child. On the other hand, patient-centred theories focus more on upholding one's own rights and respecting the rights of others.

Virtue Ethics Unlike consequentialism and deontology, virtue ethics focuses on more abstract principles. It centres around an individual's character and the virtues that define how one should be. When determining the most ethical course of action, virtue ethics raises the question of what someone who embodies certain virtues would do (Driver, 2006). Examples of such ethical virtues, as outlined by Aristotle, include justice, courage, and temperance (Kraut, 2022).

4. SOURCE for Guidance in Ethical Decision-Making—Application in Computational Machine Ethics

While keeping fundamental ethical concepts in mind, we will explore their application in the field of CME. In CME works, ethical judgments draw on various "sources for guidance" depending on the purpose and context. These "sources for guidance" are what researchers in CME rely on to inform and shape ethical judgments, ultimately influencing the ethical decision-making processes developed for machines. Researchers commonly leverage predefined rules, assessments of action consequences, and curated examples and datasets obtained from diverse sources such as experts, the general public, social media, or platforms like Amazon Mechanical Turk. The utilisation of predefined rules and the evaluation of action consequences, often backed by philosophical foundations, align with top-down decision-making approaches. Conversely, the use of datasets or examples to inform decision-making is typically associated with bottom-up decision-making approaches.

In the subsequent subsections, we will delve into sources influencing ethical decision-making, drawing upon the ethical frameworks previously examined. While these sources are often discussed individually, it is noteworthy that some scholarly work in the field combines these distinct elements to construct integrated systems guiding ethical decision-making (e.g., Van Dang et al. (2017)). Table 1 provides an overview of the current works in CME discussed in this section.

4.1 Normative Ethics—Deontology (Rule-Based)

In CME, some agents are guided by explicitly defined rules. In certain systems, a high-level set of rules, irrespective of a certain domain, is adopted, which is then applied to specific

Table 1: Overview of CME works discussed in SOURCE.

Category	Sub-Category	Example	Papers
Rule-Based	General Rules	Kant's Categorical Imperative	(Ganascia, 2007b) (Berreby, Bourgne, & Ganascia, 2017) (Lindner & Bentzen, 2018) (Singh, 2022)
		Doctrine of Double Effect	(Berreby, Bourgne, & Ganascia, 2015) (Govindarajulu & Bringsjord, 2017)
		Theory of Prima Facie Duties	(Anderson & Anderson, 2008) (Reed et al., 2016) (Anderson, Anderson, & Armen, 2005) (Svegliato, Nashed, & Zilberstein, 2021)
		Respect	(Roselló-Marín, López-Sánchez, Rodríguez, Rodriguez-Soto, & Rodríguez-Aguilar, 2022)
		Asimov's Three Laws of Robotics	(Vanderelst & Winfield, 2018)
	Domain-Specific Rules	Principle of Biomedical Ethics	(Anderson, Anderson, & Armen, 2006) (Anderson & Anderson, 2008)
		Rules of Engagement	(Arkin, 2008)
		Laws of War	(Arkin, 2008)
		Laws of Armed Conflict	(Reed et al., 2016)
		Just War theory	(Reed et al., 2016)
		Rules of the Air	(L. A. Dennis et al., 2016)
		Traffic Laws	(Thornton, Pan, Erlien, & Gerdes, 2017)
		The Highway Code	(Collenette, Dennis, & Fisher, 2022)
		Social Norms (in general)	(Carlucci, Nardi, Iocchi, & Nardi, 2015) (Malle, Scheutz, & Austerweil, 2017) (Li, Milani, Krishnamoorthy, Lewis, & Sycara, 2019)

Category	Sub-Category	Example	Papers
Consequence-Based	Utilitarianism	Act Utilitarianism	(Cloos, 2005) (Berreby et al., 2017) (Lindner, Bentzen, & Nebel, 2017) (Bourgne, Sarmiento, & Ganascia, 2021) (Limarga, Song, Pagnucco, & Rajaratnam, 2024) (Anderson et al., 2005) (Van Dang et al., 2017)
		Rule Utilitarianism	(Berreby et al., 2017)
	Balancing Benefits and Costs	Principle of Benefits vs. Costs	(Berreby et al., 2017) (Bourgne et al., 2021)
		Pareto Principle	(Lindner et al., 2017)
	Negative Consequences	Prohibiting Purely Detrimental Actions	(Berreby et al., 2017)
		Principle of Least Bad Consequence	(Ganascia, 2015) (Berreby et al., 2017) (Lindner et al., 2017) (Limarga et al., 2024)
Virtue-Based		-	(Thornton et al., 2017) (Govindarajulu, Bringsjord, Ghosh, & Sarathy, 2019) (Bench-Capon, 2020) (Svegliato et al., 2021) (Hendrycks, Burns, et al., 2021) (Vishwanath, Bøhn, Granmo, Maree, & Omlin, 2023) (Stenseke, 2023)
Datasets/ Examples-Based	From Experts		(McLaren & Ashley, 1999) (Azad-Manjiri, 2014) (Surendran et al., 2022)
	From Surveying the General Public		(Awad et al., 2018) (Noothigattu et al., 2018) (Awad, Anderson, Anderson, & Liao, 2020) (Awad et al., 2022) (Lourie, Bras, & Choi, 2021) (Forbes, Hwang, Shwartz, Sap, & Choi, 2020) (Emelin, Bras, Hwang, Forbes, & Choi, 2021) (Hendrycks, Burns, et al., 2021) (Sap et al., 2020)

Category	Sub-Category	Example	Papers
	Others		(Dwivedi, Lavania, & Modi, 2023) (Nahian, Frazier, Riedl, & Harrison, 2020)

scenarios or dilemmas for resolution. On the other hand, other systems employ domain-specific and concrete rules that are tailored to address their unique problems and are directly applicable within their respective contexts.

4.1.1 General Rules

In enabling ethical decision-making, many works use "general ethical principles" in moral philosophy or "general rules" to guide the process. Here, we define "general" ethical principles or rules as ones that follow from the philosophical ethical theories, as well as rules that are domain-independent. Concisely, rules falling under this category are those that can be applied across multiple domains.

An example of such rules is *Kant's Categorical Imperative*, which has two formulations: (1) Kant's Formula of the End in Itself and (2) the Universalisability principle. Kant's Formula of the End in Itself (Denis, 2007) emphasises the imperative to treat humanity as an end and not merely as a means to an end. The second formulation posits that one should only act according to maxims that can simultaneously be willed as universal laws. Various studies (Ganascia, 2007b), (Berreby et al., 2017), (Lindner & Bentzen, 2018) and (Singh, 2022) have actively employed and modelled the Categorical Imperative to inform ethical judgments in their respective works.

In contrast to the broad applicability of Kant's Categorical Imperative, the *Doctrine of Double Effect* applies more specifically to situations where an action is anticipated to have both positive and negative consequences (Govindarajulu & Bringsjord, 2017). A simplified version of the Doctrine of Double Effect states that, in an ethical dilemma, such an action is allowed if the harmful effects are not intended and are merely side-effects, not used to achieve benefits, and if the benefits significantly outweigh the harm. Many consider this ethical principle to be a hybrid of deontology and consequentialism. This ethical principle distinguishes itself from pure consequentialist frameworks (discussed in Section 4.2) as it does not solely focus on the consequences but also on aspects such as the actions' intentions. Berreby et al. (2015) and Govindarajulu and Bringsjord (2017) have developed frameworks that model the Doctrine of Double Effect.

Anderson and Anderson (2008) and Reed et al. (2016) both adopt Ross' *Theory of prima facie duties*. A *prima facie* duty is one which holds initial moral weight that is not absolute but flexible. The relative significance of duties can vary depending on the context, with one duty potentially outweighing another (Skelton, 2022). There are a set of fundamental ethical duties, including *beneficence*, *fidelity*, *gratitude*, *justice*, *reparation*, and *self-improvement*. These duties are also prima facie where one duty may take precedence over another depending on the circumstances. W.D. Anderson et al. (2005) is an implementation of Ross' prima facie duties where the relationship between these duties is hypothesised based on intuitions about different situations and actions are assessed under these duties. Despite its age, Anderson et al. (2005) from the early days of CME highlight the enduring relevance

of prima facie duties as more recent contributions such as Svegliato et al. (2021) who uses the idea of actions fulfilling fundamental moral duties from prima facie duties as one of the example theories in their contribution to implementing ethical theories.

Rather than focusing on a rule or a set of rules, Roselló-Marín et al. (2022) centres on enabling a single ethical value of *respect* within a survey conversational agent where the focus is on ensuring that interactions of the conversational agent with a user is respectful during a game. That is, the agent avoids disturbing the user from game-playing whilst still being able to retrieve survey responses through conversations. Specifically, the agent is punished for asking questions when the user is engaged in the game. Although this is not what one would generally consider as part of "Normative Ethics", we have chosen to include this here to highlight the possible variations as to what researchers try and aim for and get machines to adhere to in CME irrespective of their approach to implementation.

Rules for Machines. It is important to distinguish between rules for humans as opposed to machines. One classic example of rules for machines is *Asimov's Three Laws of Robotics* (Asimov, 1950) derived from a fictional novel describing general robot-human interactions. The laws state that (1) a robot must not cause harm to a human being or allow harm through inaction, (2) a robot is required to obey human orders unless such orders conflict with the first law, and (3) a robot must ensure its own existence, provided that doing so does not violate the first two laws. These laws of robotics have been used in various literature as a basis for ethical decision-making despite critiques of its flawed and fictional nature as well as the lack of scientific research to support it. Unlike the previous mentioned rules, Asimov's Three Laws of Robotics are unique in that they are one of the only ethical rules (as far as we know) to exist in CME that are intended to be exclusively applied to robots. The work by Vanderelst and Winfield (2018) is an example where a physical NAO robot has been programmed to behave according to these three laws of robotics.

AI Principles of Governments and Large Organisations. Although not yet prominent in being applied to CME, with the rapid progression and adoption of AI technologies within society, governments and other large organisations around the world have proposed principles, policies and regulations for the design and use of AI, such as Australia's 8 AI Ethics Principles, which highlights principles such as fairness, transparency, and explainability[4] as well as the recently approved EU AI Act[5]. We feel the importance of recognising the existence and relevance of these developments in this section of the survey as they may impact how ethics is incorporated into machines in the future. Our perspective is that the traditional set of "SOURCE" theories adopted within CME (summarised in Table 1) should and will eventually expand to directly encode these AI principles in some form. Other related research and discussions can be found in publications such as (Ayling & Chapman, 2022), (Langman et al., 2021), and (Almeida, Shmarko, & Lomas, 2022).

4. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles
5. https://www.technologyreview.com/2024/03/19/1089919/the-ai-act-is-done-heres-what-will-and-wont-change/

4.1.2 Domain-Specific Rules

In addition to general ethical principles, specific rules have been adopted in domains like biomedical, military, automotive, aviation, and engineering within CME. These domain-specific rules often build upon broader ethical theories. For example, *Principles of Biomedical Ethics* (Beauchamp & Childress, 1979) is a domain-specific set of prima facie duties containing a set of four duties: the Principle of Autonomy, the Principle of Beneficence, the Principle of Justice and the Principle of Non-maleficence. The Principle of Autonomy emphasises respecting individuals' rights to make their own healthcare decisions. The Principle of Beneficence involves promoting patients' well-being. The Principle of Justice is concerned with ensuring fair and equal treatment for all patients, free from bias or discrimination, while the Principle of Non-maleficence focuses on preventing harm to patients. The biomedical domain generally has more of a consensus when it comes to evaluating specific ethical dilemmas (Anderson & Anderson, 2007). Anderson et al. (2006) replaces the more general prima facie duties in (Anderson et al., 2005) with these biomedical ethical principles and Anderson and Anderson (2008) extends this work more specifically to eldercare.

Military contexts have codes of conduct like the *Rules of Engagement* established by the U.S. Military and *Laws of War* (Arkin, 2008) as well as *Laws of Armed Conflict* and *Just War Theory* (Reed et al., 2016), which inform ethical decision-making in autonomous weapon systems. In domains with less-established guidelines (possibly due to a lack of consensus), such as automotive, aviation, engineering disciplines and everyday life, studies sometimes design their own rules for CME. L. A. Dennis et al. (2016) implements and ranks rules about not damaging own aircraft and avoiding collision into hardware, people and other manned aircraft in an aviation context. Thornton et al. (2017), in an autonomous vehicle context, focuses on path tracking, vehicle occupant comfort in addition to traffic laws whilst requiring obstacle avoidance and specific vehicle slew rate limits. L. A. Dennis et al. (2016) prioritises ethics and disregards legal rules (*Rules of the Air*[6]) to be ethical. This highlights the interplay between ethics and law. Collenette et al. (2022) examines the UK's legal rules for autonomous vehicles (*The Highway Code*[7]), suggesting that considering legal rules may be a part of ethical reasoning. This underscores the potential necessity of legal rules within ethical decision-making processes.

In addition to legal rules, there have also been considerations of domain-independent social norms in social robots (Carlucci et al., 2015; Malle et al., 2017) such as greetings and expressions of gratitude when interacting with people; shaking hands when you meet someone and facing the person when you speak to them. Li et al. (2019) further examine scenario-specific (e.g., ordering groceries) and context-specific (e.g., entering bathroom with a bursting pipe) norms for domestic robots relating to aspects including safety, being considerate and privacy.

Although ethical principles, social norms and legal rules serve distinct purposes, they can overlap, and the distinction between these concepts can be ambiguous and is evolving within CME. The ambiguity is particularly evident between ethical principles and social norms, both of which are heavily context-dependent. Others (Collenette et al., 2022) have also drawn connections between ethical principles and legal rules as aforementioned. Hence,

6. https://publicapps.caa.co.uk/docs/33/CAP393REFERENCE_ONLY.pdf
7. https://www.gov.uk/guidance/the-highway-code

we have included discussions about all three concepts and do not rigorously differentiate between them. Additionally, Fox and Rey (2024) distinguish between two types of "ethical requirements": (1) locally simple requirements, exemplified by road traffic regulations which are more universal, as opposed to (2) locally complex requirements where requirements are less well-defined and may be subjective.

4.2 Normative Ethics—Consequentialism (Consequence-Based)

Many of the contributions in the field take on a consequentialist approach in making decisions where actions are deemed ethically permissible by examining the consequences of the actions. Often, these consequences are assigned a utility value to represent how good or bad the action is to enable comparison and evaluation to come to an ethical decision. Due to its "modifiability", taking on a consequentialist approach has been a popular, if not the most popular, choice in CME research. Although fewer variations of consequentialism are observed within the field compared to rule-based deontological approaches, there are still a few forms of consequentialism in CME. They are divided into three themes and outlined below.

4.2.1 Utilitarianism—A Focus on Utility

Utilitarianism stands out as one of the most influential forms of consequentialism in CME. *Act Utilitarianism*, a form of utilitarianism, asserts that the most permissible action is the one leading to the best overall outcome for the majority involved. Numerous authors in CME (Cloos, 2005; Berreby et al., 2017; Lindner et al., 2017; Bourgne et al., 2021; Limarga et al., 2024), have embraced this idea. Some studies leave the definition of utility open and do not exemplify/specify how such values are determined (Berreby et al., 2017). A number of works define utilities that represent aspects relating to individuals' health (Lindner et al., 2017; Cloos, 2005; Bourgne et al., 2021). As an example, act utilitarianism, among numerous ethical theories Lindner et al. (2017) implements, assigns utilities specifically reflecting the number of deaths as a consequence of extreme dilemmas (i.e., life or death situations). Many works exemplify their utilitarian approach using such extreme ethical situations to highlight their approach and prevent ambiguities. Though, where death or individuals' health is less relevant, for example in a scenario involving lying, utility is used to represent the severity of the consequences or the benefits which the act brings about (Lindner et al., 2017).

In many of the above studies, assessing action consequences is simply a matter of summing up the utilities of all consequences of a given action for all affected parties given one action. On the other hand, Limarga et al. (2024), although leaving the definition of how utilities are defined open (i.e., the "goodness" value), provides various definitions of aggregation functions. Given a plan of action (instead of a single action) and a situation, an aggregation function can provide a "goodness" value for the plan. It draws a connection between strategies in game theory and such aggregation functions.

Variations of Act Utilitarianism are also observed in CME. Anderson et al. (2005) implements a version of Act Utilitarianism (*Jeremy*), known as *Hedonistic Act Utilitarianism*, where the primary focus lies on maximising the "pleasure" experienced by the impacted parties. Different to Jeremy, Act Utilitarianism is combined with deontological rules in the work by Van Dang et al. (2017), where utility calculations find their foundation in Asimov's Laws of Robotics.

In contrast to Act Utilitarianism, *Rule Utilitarianism* provides a broader societal perspective as it assesses the permissibility of an action by considering its impact if all members of society were to follow the same rule in similar situations. This is explored by Berreby et al. (2017) in their work where, rather than choosing an action that produces the greatest utility, the "moral rule" (that will be applied to everyone) which produces the greatest utility is chosen. An example is given by the author using the rules of "Do not steal", where stealing is always impermissible even if you are stealing food to feed a starving child. However, this may become permissible if evaluated under Act Utilitarianism.

4.2.2 Balancing Benefits and Costs

Finding a balance between benefits and costs from actions is observed in some consequentialist theories used in CME. These principles delve into the complexities of ethical decision-making when faced with situations where possible actions potentially yield some degree of negative consequence alongside positive outcomes. The *Principle of Benefits vs. Costs*, adopted by studies such as (Berreby et al., 2017; Bourgne et al., 2021), determine that actions are only permissible if the good consequences outweigh the bad consequences. Alternatively, we see the use of the *Pareto Principle* where an action is permissible if there is no other action with more good consequences or less bad consequences in the work by Lindner et al. (2017).

4.2.3 Prevention and Minimisation of Negative Consequences

Rather than looking at the benefits of actions at all, there are consequentialist ethical principles that specifically focus on preventing and minimising negative consequences of actions. One such principle is *Prohibiting Purely Detrimental Actions* which asserts that actions with only bad consequences are impermissible. This approach raises questions about the permissibility of actions when their outcomes are exclusively negative. Nevertheless, the principle is modelled by Berreby et al. (2017) using logical predicates.

In situations where all available actions lead to negative consequences to varying extents, the *Principle of Least Bad Consequence* comes into play. This principle suggests that, when faced with a set of undesirable outcomes, the ethically right choice is the one associated with the least harmful consequence. It introduces a nuanced approach to ethical decision-making, acknowledging the imperfection of available options. (Ganascia, 2015; Berreby et al., 2017; Lindner et al., 2017; Limarga et al., 2024) apply the principle in their approaches to CME. By considering these ethical theories, machines may navigate the complexities of ethical decision-making in scenarios where purely positive outcomes are unattainable.

4.3 Normative Ethics—Virtue Ethics (Virtue-Based)

Determining ethical choices based on a defined set of rules or assessing action consequences are common practices used by researchers for their methodology on ethical theories. In recent years, we have seen a rise in implementations of artificial virtuous agents such as (Thornton et al., 2017; Govindarajulu, Bringsjord, Ghosh, & Sarathy, 2019; Bench-Capon, 2020; Svegliato et al., 2021; Vishwanath et al., 2023; Stenseke, 2023) utilising ideas from Virtue Ethics, which are relatively more abstract and deemed as "uncodifiable" and hence more challenging to implement on machines. We understand this as a contributing factor to the discrepancies in the interpretation of virtue ethics within CME.

As aforementioned, virtue ethics assesses the ethical nature of an action by considering whether a virtuous person, acting in accordance with their character, would perform such an action in a given situation. One ethical framework presented by Svegliato et al. (2021) is founded upon this idea where actions are chosen if they follow any "moral trajectory" performed by a *moral exemplar*, that is, a human with the relevant domain expertise (who would be considered the virtuous person). The idea of following a moral exemplar is also discussed in the work by Hindocha and Badea (2022), where the authors argue the importance of clinicians in healthcare as moral exemplars for virtuous machines. Following from this idea of imitating the behaviour of virtuous agents, Hendrycks, Burns, et al. (2021) attempts to incorporate Virtue Ethics into its benchmark ETHICS[8] dataset, designed for language models trained to identify character traits either present or absent within given scenarios. More general examples of datasets will be discussed in the subsequent section. Vishwanath et al. (2023) highlights an important advantage of virtue ethics being the trait of life-long learning where a virtuous individual will try to improve their decisions in future similar situations. This idea is further supported by (Stenseke, 2023), which extensively explores features of virtue ethics and takes on a similar interpretation.

In another way, the theory focusing on one's character inspires the inclusion of "role morality" in Thornton et al. (2017)'s ethical decision-making process in automated vehicle control. "Role morality" is the idea that the permissibility of certain behaviour is dependent on the role of the individual or context. The paper introduces the idea of "Vehicle Character", where the role of a vehicle determines the application strength of certain rules on a vehicle and the cost of violation. For example, it is acceptable for an ambulance carrying a patient with a life-threatening condition to run a red light due to its societal role. Govindarajulu, Bringsjord, Ghosh, and Sarathy (2019) formalises a version of Zagzebski's idea of virtue ethics; *Exemplarist Virtue Theory* (Zagzebski, 2010) where "exemplars" are identified by the emergence of the emotion of admiration upon encountering these exemplars. These exemplars are then studied to understand what traits they possess, forming the Exemplarist Virtue Theory. Interpreting and applying virtue ethics differently, virtues are viewed by Bench-Capon (2020) as preferences between values or needs, defining the notions of being *selfish* (unethical), *altruistic* (ethical) and *sacrificial* (supererogatory), which are implemented in its virtuous system.

4.4 Descriptive Ethics—Datasets/Examples-Based

While rules, consequences, and virtues provide foundational guidance in ethical decision-making, their inherent abstraction often poses challenges in practical application to real-world scenarios. Bridging this gap requires substantial human effort. Additionally, nuanced situations or ethical dilemmas may emerge that are not adequately addressed by high-level guidance, leading to unforeseen consequences upon evaluation. In such cases, consulting a human expert becomes essential for sound judgment. Alternatively, leveraging datasets containing ethical examples, combined with machine learning, proves pivotal in enhancing machine decision-making capabilities, especially in addressing novel ethical dilemmas. Recent advancements in the field emphasise the curation and utilisation of such datasets. This type

8. Note: All dataset names referenced in this paper are presented in uppercase, following the common practice in the literature.

of approach, however, introduces challenges related to transparency as the decision-making process becomes dependent on datasets.

4.4.1 From Experts

Early research in CME employed various approaches, including the utilisation of a limited set of manually analysed example cases to guide the process of generating ethical suggestions or conclusions. For instance, McLaren and Ashley (1999) conducted a study in the domain of engineering ethics, where example cases were employed to evaluate novel situations. The National Society of Professional Engineers Board of Ethical Review (NSPE BER) analysed and published 400 ethical scenarios that professional engineers commonly encounter in their work. These scenarios covered a wide range of themes, such as Gifts, Plagiarism, Criticism of Another Engineer, Disclosure of Potential Conflicts of Interest, and more. The publication also included pertinent abstracts of the code of ethics and corresponding recommendations for each situation. While McLaren and Ashley (1999) relied on available examples, these domain-specific example cases are not readily available.

In more recent years, the widespread adoption of machine learning has led to a surge in the creation of various datasets as opposed to the few example cases from the aforementioned work by McLaren and Ashley (1999). Azad-Manjiri (2014); Surendran et al. (2022) have taken a different approach by creating their own datasets through the use of questionnaires in the relevant domains. Azad-Manjiri (2014) gathered responses from biomedical ethicists regarding ethical situations within the biomedical context. In contrast, Surendran et al. (2022) conducted a survey investigating how experts and individuals (referred to as "folks") approach ethical decisions in healthcare and game playing scenarios. The "folks" responded with Yes or No to indicate the appropriateness of actions, while experts were asked to apply different ethical theories to the same situations.

4.4.2 From Surveying the General Public

Most existing datasets primarily comprise judgments from the general population. One common approach to collecting such examples is through surveys conducted among the general public. *The Moral Machine experiment*, described in (Awad et al., 2018), utilised an online platform to gather over 40 million ethical decisions on scenarios involving autonomous vehicles. The experiment spanned ten different languages and included participants from various regions worldwide. The platform presented users with dilemmas, often variations of the Trolley Problem, where accidents were unavoidable. Users were asked to choose one of two possible decisions based on the generated dilemmas, which were influenced by nine factors, such as choosing between saving humans or pets, staying on course or swerving, protecting passengers or pedestrians, and prioritising the young or the elderly. Users had the option to provide background information, which was used to identify patterns in ethical preferences among different groups and regions. Examples that utilise this collected data include (Noothigattu et al., 2018; Awad et al., 2020). Noothigattu et al. (2018) takes the stance where societal choice acts as a fallback when principle-based approaches do not lead to a decision. Another study was conducted by Awad et al. (2022), which surveyed 400 participants via Amazon Mechanical Turk service regarding the notion of rule-breaking. The

study explored the acceptability of cutting the line in three contexts: delicatessen, bathroom, and airport.

Researchers have also leveraged diverse large-scale text datasets to explore neural language models' capabilities in ethical judgment. Notable datasets include *SCRUPLES* (Lourie et al., 2021), featuring 32k real-life ethical situations, and *SOCIAL-CHEM-101* (Forbes et al., 2020), offering real-life scenarios and Rules of Thumb from subreddits[9], *ROCStories* (Mostafazadeh et al., 2016), and *Dear Abby*[10]. The *MORAL STORIES* corpus (Emelin et al., 2021) comprises 12k crowd-sourced stories depicting ethical reasoning in daily life. *ETHICS* (Hendrycks, Burns, et al., 2021) integrates well-established ethical theories, emphasising clarity and removing scenarios with low worker agreement rates. The *SOCIAL BIAS INFERENCE CORPUS* (Sap et al., 2020) focuses on ethical situations related to social biases from various social media platforms. These datasets, sourced globally (primarily from English speakers from English-speaking countries), are reflective of descriptive ethics, aligning with human values and decision-making processes.

4.4.3 Others

EtiCor (Dwivedi et al., 2023) is a text-based corpus designed for large language models consisting of social norms from five different regions around the world. It is interesting to note that this differs from previous datasets consisting of example situations, the dataset focuses on sentences describing social norms. Additionally, rather than gathering expert or public opinions, the data of this corpus comes from collecting texts from a variety of publicly available authentic sources such as websites, tour guide pamphlets and magazines.

Unlike all previous examples, Nahian et al. (2020) presents a highly unique approach by training machine learning models to classify situations using a rather unconventional source: the children's educational comic strip, *Goofus and Gallant*. The authors argue that these stories are useful in learning a "strong and robust prior for value alignment" as the characters embody normative and non-normative behaviours within situations. These situations are presented through text-based descriptions, encoding societal norms and reflecting human values.

5. Ethical DECISION-Making Process

This section categorises various techniques in CME into three main groups: Top-down, Bottom-up, and Hybrid approaches, following the high-level classification proposed by Allen, Smit, and Wallach (2005). *Top-down* approaches are guided by established ethical theories, shaping the behaviour and decision-making processes within a system. This may involve implementing ethical theories or specific rules to derive conclusions in various scenarios. In contrast, *bottom-up* approaches rely on machines learning and imitating ethical behaviour from examples of human behaviour. Allen compares bottom-up approaches to how a young child learns ethics in a social context, identifying appropriate behaviour without explicit theory. *Hybrid* approaches integrate elements of both top-down and bottom-up strategies. The categorisation of techniques can be ambiguous, as some authors may view particular

9. https://www.reddit.com/
10. https://www.uexpress.com/life/dearabby/archives

techniques as top-down and others as bottom-up, both with reasonable justifications. For the purpose of this survey, we conduct the categorisation based on our interpretation and definitions defined in this survey. Table 2 provides an overview of how we have categorised the works discussed in this section.

Table 2: Overview of CME works discussed in DECISION.

Category	**Sub-Category**	**Papers**
Top-Down	Logic-Based	(Bringsjord & Taylor, 2012) (Grandi, Lorini, Parker, & Alami, 2023) (Berreby et al., 2017) (Govindarajulu & Bringsjord, 2017) (Govindarajulu, Bringsjord, Ghosh, & Peveler, 2019) (Govindarajulu, Bringsjord, Ghosh, & Sarathy, 2019) (Horty, 2001) (Limarga, Pagnucco, Song, & Nayak, 2020) (Pagnucco, Rajaratnam, Limarga, Nayak, & Song, 2021) (Bonnemains, Saurel, & Tessier, 2018) (Ganascia, 2007a) (Hooker & Kim, 2018) (Pereira & Lopes, 2007) (Pereira & Saptawijaya, 2009) (Han, Saptawijaya, & Moniz Pereira, 2012) (L. A. Dennis et al., 2016) (Tufis & Ganascia, 2015) (Wiegel & van den Berg, 2009) (Neto, Silva, & Lucena, 2011) (Cointe, Bonnet, & Boissier, 2016) (Cranefield & Dignum, 2020) (Cranefield, Winikoff, Dignum, & Dignum, 2017) (Bremner, Dennis, Fisher, & Winfield, 2019) (Honarvar & Ghasem-Aghaee, 2009) (Neto, Silva, & Lucena, 2010)
	Utility-Based	(Anderson et al., 2005) (Van Dang et al., 2017) (Winfield, Blum, & Liu, 2014) (Vanderelst & Winfield, 2018) (Lindner, Mattmüller, & Nebel, 2020)
	Others	(Loreggia, Mattei, Rossi, & Venable, 2018) (Bendel, 2016)
Bottom-Up	Case-Based Reasoning	(Ashley & McLaren, 1994) (McLaren, 2003)
	Supervised Learning	(Guarini, 2006) (Howard & Muntean, 2017) (Jiang et al., 2021) (Noothigattu et al., 2018)

Category	Sub-Category	Papers
	Reinforcement Learning	(Krening, 2023) (Vishwanath et al., 2023) (Abel, MacGlashan, & Littman, 2016) (Wu & Lin, 2018) (Berberich & Diepold, 2018) (Roselló-Marín et al., 2022)
Hybrid	Inductive Logic Programming	(Anderson et al., 2006) (Dyoub, Costantini, & Lisi, 2019) (Awad et al., 2020)
	Others	(Madl & Franklin, 2015) (Azad-Manjiri, 2014) (Honarvar & Ghasem-Aghaee, 2009) (Fox & Rey, 2024) (Ramanayake & Nallur, 2024)

5.1 Top-Down Approaches

Top-down approaches to ethical decision-making are founded on the basis of some existing ethical theory or principles (discussed in Section 4). In this section, we will sub-categorise approaches based on the model/representation on which decisions are based. There are three main types of top-down approaches, namely, logic-based, utility-based and other approaches. We have included an "Others" section to discuss some studies that have yet to form a distinct cluster of works within CME.

5.1.1 Logic-Based

In the exploration of top-down approaches in CME, the adoption of logic-based technologies stand out as a prevalent methodology. This involves the translation of specific sets of ethical principles into a formal representation. Subsequently, an automated reasoning process is initiated, aiming to determine the most ethically permissible course of action (involving a certain type of reasoning and the technology used to automatically execute such reasoning). Logic-based approaches are renowned for their transparent and well-defined nature, contributing significantly to explainability and traceability—attributes crucial in the ethical domain where determining the most ethical choice may be ambiguous and prone to controversy. However, a commonly recognised challenge with this type of approach lies in its implementation within real-life ethical situations. As a consequence, some logic-based CME contributions only describe or partially implement their ideas (e.g., provide a logic formalisation).

Bringsjord and Taylor (2012) takes on an approach based on divine-command ethics to regulate behaviours of a war-fighting robot where commands as to how to behave come directly from God. It introduces the divine-command logic, their variation of *deductive logic*, LRT^* (based on a paper-and-pencil divine-command logic LRT[11] (Quinn, 1978)) to formalise and implement ethical rules which allows for the automation of checking logical

11. "the theological version of the logic of requirement".

inferences. They ultimately aim to ethically regulate robot behaviours using LRT* which in this work enables the automation of proof checking but not proof discovery. The authors highlight that proof discovery is necessary to respond to queries such as "Is it permissible for me to destroy this building?". Following this, temporal logic extends deductive logic and enables the representation and reasoning that involve the notion of time or order. Specifically, Grandi et al. (2023) uses *linear temporal logic* to express an agent's values and goals. Plans of action are evaluated against these values and goals through lexicographic preference modelling, where values are lexicographically ordered based on importance to help select the most ethical plan.

With temporal considerations in mind, *Event Calculus*, allowing for the representation and reasoning of specific events and actions over time, permits the modelling of real-life situations that change dynamically. The work by (Berreby et al., 2017), implemented using answer set programming, utilises event calculus to depict various ethical scenarios and principles within consequentialist and deontological ethics. Combining the concepts of deontic logic (the logic of obligation) and event calculus, Govindarajulu and Bringsjord (2017); Govindarajulu, Bringsjord, Ghosh, and Peveler (2019); Govindarajulu, Bringsjord, Ghosh, and Sarathy (2019) use *Deontic Cognitive Event Calculus* to formalise the Doctrine of Double Effect and Virtue Ethics. Horty (2001) proposes a framework that extends *Deontic Logic* to represent and reason about what agents ought to do. The author draws parallels between making a choice under uncertainty in decision theory and taking action in indeterministic time. A preference ordering is adopted to define the best actions an agent should take given a situation and to determine the outcomes an agent should aim to ensure.

Similarly, Limarga et al. (2020) proposes an implementation model that considers the need to reason over a sequence of actions in real-life situations. The proposed implementations involve non-monotonic reasoning performed on formulations of consequentialism and deontology using *Situation Calculus* to represent knowledge through actions, fluents, and situations. Non-monotonic reasoning allows for the revision of beliefs and conclusions in light of new information. Instead of non-monotonic reasoning, Pagnucco et al. (2021) combines situation calculus with epistemic reasoning for decision-making grounded in consequentialist and deontological ethics. Bonnemains et al. (2018) introduces a novel approach to modelling ethics akin to both event calculus and situation calculus. Ganascia (2007a) formalises ethical theories of Aristotelian consequentialism, Kantian Ethics (deontology) and Constant's theory of Principles using *non-monotonic logic*, specifically making use of answer-set programming to provide an implementation.

Hooker and Kim (2018) formulates existing ethical theories, namely (1) the Generalization, (2) Joint Autonomy and (3) Utilitarian principles into *quantified modal logic*, enabling more precise moral principles for computation. The authors believe that applying ethical principles to particular situations should remain a human's task. Specifically, the programmer, involving their knowledge and beliefs. *Abductive logic* allows for the inference of information based on known facts and is adopted by Pereira and Lopes (2007); Pereira and Saptawijaya (2009) to prospectively examine consequences of hypothetical moral judgments and is also in Han et al. (2012)'s work for reasoning under uncertainty. These works have a focus on creating and implementing prospective logic agents which actively anticipate future events, preemptively adapt and conduct moral reasoning considering these events under uncertainty.

Several approaches (L. A. Dennis et al., 2016; Tufis & Ganascia, 2015; Wiegel & van den Berg, 2009; Neto et al., 2011; Cointe et al., 2016; Cranefield & Dignum, 2020; Cranefield et al., 2017; Bremner et al., 2019) have adopted a *Belief-Desire-Intention (BDI)* model (Rao & Georgeff, 1995) as a basis and incorporate an additional ethical component to the model. As the name suggests, this model is about enabling agents to reason about their beliefs (knowledge of the world), desires (goals and preferences) and intentions (actions) to come to a decision. Wiegel and van den Berg (2009) follows deontological ethics and embeds a modal logic framework DEAL (Deontic Epistemic Action Logic) into its BDI model. L. A. Dennis et al. (2016) takes the ethical decision-making process in a BDI agent beyond just decision-making and investigates the formal verification of choices. The authors propose to prove decisions using model checking on formally specified properties, examining all possible executions for a given scenario. Instead of solely considering whether actions are moral or immoral and should or should not be taken, Neto et al. (2011) introduces the Norm-Belief-Desire-Intention (NBDI) model, based on the representation of norms from Neto et al. (2010). Implemented on the Normative Jason platform (Rafael H. Bordini, 2007), this model focuses on several key functions: (1) determining whether norms should be adopted according to an agent's desires and intentions; (2) evaluating the advantages and disadvantages of fulfilling or violating the norm; (3) checking and resolving conflicts between adopted norms; and, (4) ultimately selecting desires and plans based on whether the agent opts to fulfill the norm.

While logic-based approaches in CME offer a promising avenue for formalising ethical principles and enhancing our understanding of ethical considerations, their implementation presents notable challenges. The removal of ambiguities and the promotion of clarity are central strengths acknowledged by many researchers. However, the translation of these principles into fully realised systems remains an ongoing endeavour, with limited examples of comprehensive implementations in real-life scenarios. The inherent inflexibility and the imperative to encode all relevant information within formal representations pose significant hurdles as currently this can only be done manually. These approaches thrive in well-defined environments with specific assumptions but may encounter limitations in addressing the dynamic and complex nature of real-life ethical scenarios.

5.1.2 Utility-Based

A cluster of work within CME select an ethical action based on an assigned or calculated value founded on the basis of some ethical theory or principle. Utility-based approaches focus on using numerical values for evaluating ethics and tend to provide a more easily implemented prototype as opposed to other types of top-down methodologies. In the earlier days of CME, Anderson et al. (2005) proposed *Jeremy*, a relatively straight-forward implementation of Hedonistic Act Utilitarianism, where pleasure and displeasure of individuals are considered. It is a system that receives user input on information such as the action, individual affected, rough estimate of the amount and likelihood of pleasure/displeasure if the action was taken (this is done for each action and all persons affected by the action) and presents the user with the action(s) where the net pleasure is the greatest.

Some studies (Van Dang et al., 2017), (Winfield et al., 2014) and (Vanderelst & Winfield, 2018) assess anticipated consequences of different actions to select the most ethical action

specifically based on Asimov's Three Laws of Robotics (Asimov, 1950). Van Dang et al. (2017) implements a home service robot that would evaluate each possible action according to the three laws and select one to perform based on an assigned utility score. A value between -5 and 5 is assigned to each possible action based on the consequences it leads to for each of the three laws. These values are then summed up, and the action with the highest value is chosen. The paper uses ordering food as an example, where the agent decides whether or not to follow a food order based on the current state of the individual's health and the health consequences of the food ordered. To do so, the agent refers to a table containing information about the family and healthcare, which is stored in the agent's semantic memory (SMem).

Other studies (Vanderelst & Winfield, 2018) and (Winfield et al., 2014) take on similar approaches in providing alternative implementations of Asimov's Three Laws of Robotics. Both use some kind of simulation technique to simulate the consequences of actions, which are then evaluated using utility scores like (Van Dang et al., 2017). In (Vanderelst & Winfield, 2018), the Robot Controller provides a set of potential actions to take, which gets fed into the Ethical Layer of the system. Within the Ethical Layer, the Simulation Module is initialised with the state of the world, the robot, and the human involved. The Evaluation module evaluates each action based on the simulated consequences for both the human and the robots. It then combines this information into a single metric to indicate the desirability of an action. This value is based on factors related to Asimov's three laws of robotics, such as the safety level of the human and whether the robot had executed an order given by the human. The most desirable action is taken by the robot via the Robot Controller.

Winfield et al. (2014) has a Consequence Engine that loops through all possible actions in each scenario and simulates the consequences. These consequences are then evaluated by the Action Evaluator, where each outcome is assigned a numerical value representing the estimated degree of danger (0 means Safe and 10 means Fatal). The system has a separate Safety/Ethical Logic (SEL) layer compared to (Vanderelst & Winfield, 2018). The SEL involves logic, which comes into play when the most ethical action is taken (that is, the action(s) with the least unsafe human outcome(s)) when not all robot actions lead to the human being equally safe. Van Dang et al. (2017) only applies to simple/straight-forward scenarios where anticipated consequences of an action are stated prior to any decision-making, whereas other approaches (Vanderelst & Winfield, 2018; Winfield et al., 2014) use simulations to predict consequences for potentially more complex scenarios.

Rather than looking at the permissibility of individual actions, Lindner et al. (2020) focuses on analysing moral permissibility in the context of sequences of actions (i.e., entire plans) according to different ethical principles which are formalised. The approach assumes the use of utility functions that output utility values to describe actions and facts but do not define how such values are determined for an action.

5.1.3 Others

Instead of representing information using logic or numerical utility values, Loreggia et al. (2018) takes a different method of formalisation by constructing two Conditional Preference networks (CP-nets), one representing individual ethical preferences and the other encoding ethical principles. CP-nets were introduced by (Boutilier, Brafman, Domshlak, Hoos, &

Poole, 2004) to capture preference relations through a graphical probabilistic representation. The individual preferences are evaluated to determine whether they are "close enough." to the ethical principles leading to an ethical action using approximation algorithms. Bendel (2016) utilises a more straightforward graphical representation of decision trees to model ethical decision-making processes for a simple moral machine. The authors describe a manual process in developing a decision tree that captures the ethics of a particular activity in a specific domain. Their work is exemplified using the activity of driving. The tree consists of decision nodes based on the situation starting from a root node (e.g., anything less than 40m ahead on the road) branching out and eventually coming to a decision for an action to take (e.g., emergency braking). The simplicity of such representation enables clarity and accessibility in understanding the decision process.

5.2 Bottom-Up Approaches

Bottom-up approaches learn from examples of ethics and are mostly characterised by machine learning-based approaches. Machine learning-based approaches vary in their program model (e.g., decision tree, neural network) and the algorithms chosen to navigate this space to come to some conclusion. In this section, we will broadly organise the CME literature of bottom up approaches based on the type of algorithms they adopt, namely supervised learning, reinforcement learning and case-based reasoning (a non-learning but example-based methodology). Purely using learning-based methodologies in coming to ethical conclusions can be dangerous given their opaque nature and the difficulty in explaining and justifying how these conclusions are made. However, they can be seen as more practical and flexible in terms of implementation in the real world as they are able to come to an ethical conclusion in novel situations, which is often a challenge with other approaches discussed in this paper.

5.2.1 Case-Based Reasoning

Some of the earlier works in CME have taken on a case-based approach such as the Truth-Teller (Ashley & McLaren, 1994) and SIROCCO (McLaren, 2003). Case-based reasoning is the process where problems are evaluated by comparisons to existing cases. McLaren et al. (Ashley & McLaren, 1994; McLaren, 2003) created these two ethical assistants that provided relevant information to humans for difficult ethical situations. Neither of these programs comes to any ethical conclusions but rather simply synthesises provided and existing information from an ethical perspective.

Truth-Teller (Ashley & McLaren, 1994) compares a pair of ethical dilemmas (each represented as a network of actors, relationships, actions performed and reasons for their actions) to assist humans in determining whether one should tell the truth in these dilemmas. The program compares the pair of dilemmas by mapping reasons for actions and highlights similarities and differences between them. A human-readable comparison output is presented, highlighting reasons for both disclosing and withholding the truth that apply to both dilemmas. Additionally, Truth-Teller identifies reasons that apply more strongly to one case or exclusively to one of the two cases, which is useful for generating ethical arguments. The approach was considered moderately successful at making comparisons between ethical dilemmas when compared to student-written dilemma comparisons rated by professional ethicists.

Following Truth-Teller, McLaren created SIROCCO (McLaren, 2003), which is based on interpretive case-based reasoning (Kolodner, 1993) where evaluations of situations are conducted in the context of previous experience. It combines general abstract ethical principles to concrete existing past cases in reasoning and making decisions about ethical situations. Unlike Truth-teller, SIROCCO focuses on engineering ethics principles and considers only one ethical case and the associated ethical question. The ethical case is expressed in the Engineering Transcription Language (ETL), which provides a structured and simplified representation. This language is designed to facilitate reasoning and highlights the chronological sequence of events within ethical scenarios. Using a graph-mapping algorithm, SIROCCO retrieves and outputs a list of potentially relevant ethic codes, past cases, and any additional suggestions. SIROCCO does not match a human reasoner, but it was much more accurate in retrieving codes and cases than most other methods at the time.

Case-based reasoning is a noticeably outdated approach within CME and we believe this is due to the rise of machine learning techniques which have dominated the AI field in recent years. This is especially the case as both case-based reasoning and machine learning stem from a similar goal of making decisions based on examples where machine learning appears to be more scalable and efficient. In the following sections, we will examine some of the machine learning approaches in CME.

5.2.2 Supervised Learning

Supervised learning, a widely favoured paradigm in machine learning, has naturally emerged as a direction of research in CME, aligning with the growing prominence of the machine learning domain. Dependent on labelled training data and training examples to infer functions ultimately used to make decisions (Sarker, 2021), supervised learning was one of the first kinds of learning approaches taken on by CME researchers. It tends to lend itself to ethical tasks involving the classification of ethically relevant cases as ethically acceptable or unacceptable.

An early example of such (Guarini, 2006) follows the particularist viewpoint. Rather than adhering to one set of predefined ethical principles for ethical agents, the work uses a neural network model to classify the acceptability of ethical cases involving scenarios where individuals are either killed or left to die. It specifically considers a meta-network on top of a single recurrent network for single case classification. The outputs of the single recurrent network act as the input for the Metanet that attempts to identify similar case pairs that exhibit a contrast in their classification from the single recurrent network and hence flag specific features that make a difference in the ethical classification.

Howard and Muntean (2017) followed Guarini's exploration of a learning-based approach using neural networks. They endorse particularism but are also accepting of ethical principles discovered through learning (although not central to its design). They combine neural networks with evolutionary computation, where neural networks are trained to learn human ethical decision patterns and generalise "robo-virtues" from data. Evolutionary computation, a type of optimisation algorithm inspired by the natural evolution process, is used to find the best individual neural network from a population of networks. The authors implement this idea with a lifeboat scenario (Hardin, 1974). At the time of Howard and Muntean's

contribution (Howard & Muntean, 2017), Guarini's works were still some of the few concrete implementations of bottom-up learning-based "Artificial Autonomous Moral Agents".

A relatively recent and controversial contribution (Jiang et al., 2021) builds Delphi, a prototype model that follows the deep learning approach on a much larger scale with the ability to perform multiple types of reasoning tasks (i.e., free-form Question Answering, yes/no Question Answering and relative Question Answering) and come to an ethical conclusion on a wide variety of everyday situations. In training Delphi, the authors also provide a "moral textbook customised for machines", that is a large-scale corpus compiled from existing datasets (see Section 4.4) to educate machines about human commonsense moral reasoning. The authors were able to achieve a high accuracy with its model relative to human ethical decision-making. This particular contribution has created significant interest within CME and sparked critiques (Talat et al., 2022), analysis (Fraser, Kiritchenko, & Balkir, 2022) as well as enabled various follow-up studies based on Delphi's contributions (Ammanabrolu, Jiang, Sap, Hajishirzi, & Choi, 2022; Pyatkin et al., 2023).

Talat et al. (2022) criticises Delphi in several aspects. They raise concerns about the data used for the model, questioning what constitutes a good representative sample of situations for the model to learn from. Additionally, they argue that foreseeing a normative application of a model based on descriptive ethics, as proposed by Jiang et al. (2021), is problematic. They also argue that moral questions should not be treated merely as tests to be passed. Furthermore, automating ethical decisions in this way is seen as problematic because it inhibits debate and contestation. Delphi has also received some criticism from the general public as people test its output in various scenarios.[12, 13] Criticisms of contentious, inconsistent, illogical, and offensive responses surround the outputs that Delphi provides to specific input situations. Additionally, the way the input is worded can be tweaked to achieve certain outputs, due to linguistic nuances and complexities.

Most purely bottom-up supervised learning approaches are based on some neural network model. Bostrom and Yudkowsky (2014) expresses that decision tree or Bayesian network type machine learners are much more transparent in understanding their inner-workings relative to algorithms based on complicated neural networks which often operate as "black boxes". The internal workings of neural network models do not provide straightforward insights as to how or why decisions are made. In ethical contexts, as AI algorithms play an increasingly large role in society, being "transparent to inspection" by others is important. Few decision tree approaches exist within the field. However, rather than being purely bottom-up approaches, we identify them as top-down (Bendel, 2016) and hybrid (Azad-Manjiri, 2014) approaches (discussed in the respective sections of this survey).

Many of the above learning-based approaches focus on learning from examples of ethical scenarios where the ethical choice is clear. Noothigattu et al. (2018) takes an approach inspired by social choice as they focus on aggregating people's opinions on ethical dilemmas to make ethical decisions by combining learnt models of individual preferences on different dilemmas into a single model representing the collective preference, specifically from scenarios and data collected from the Moral Machine (Awad et al., 2018). Different from previous neural network-based approaches, preferences are modelled using what is referred to as "permutation processes".

12. https://www.nytimes.com/2021/11/19/technology/can-a-machine-learn-morality.html
13. https://www.theverge.com/2021/10/20/22734215/ai-ask-delphi-moral-ethical-judgement-demo

5.2.3 Reinforcement Learning

There is a cluster of work in CME specifically on reinforcement learning, where the central idea is that an agent receives rewards or punishments based on its actions to help it learn what to do in future scenarios. Krening (2023) is an example of work from reinforcement learning where the authors draw parallels between Q-learning (a specific model-free reinforcement learning algorithm) and Utilitarianism, where they both are about choosing actions that lead to optimal consequences. It promotes the idea of a human-machine team to make ethical decisions where the human defines the high-level information and the machine performs the actual calculations. Specifically, the human decides the possible actions and consequences for the states and assigns values to the consequences based on how good or severe they are. Based on this information, the machine learns an optimal mathematical function to represent Utilitarianism by calculating and maximising utilities of different sequences of actions. The system outputs a chosen action and additionally an ordering on the possible actions in its efforts to improve explainability of the system. Combining reinforcement learning with a deep neural network model, Vishwanath et al. (2023) implements an artificial virtuous agent in role-playing games where the character is required to make decisions in ethical dilemmas. Other works in reinforcement learning include (Abel et al., 2016; Wu & Lin, 2018; Berberich & Diepold, 2018; Roselló-Marín et al., 2022).

There are a range of reinforcement learning approaches in CME, differing across multiple dimensions. Studies that implement ethical decision-making within reinforcement learning agents vary in the algorithm/model adopted, such as the aforementioned use of Q-learning (Krening, 2023), the incorporation of neural networks (Vishwanath et al., 2023) and the use of Markov Decision Processes (MDPs), specifically partially observable Markov Decision Processes (POMDPs) in the work by Abel et al. (2016). They also vary in the way in which rewards and punishments are implemented. Abel et al. (2016) utilises a single reward function based on a previous state and an action taken whilst Wu and Lin (2018) uses reward shaping to provide additional rewards for an agent. The reward is given for ethics-related actions identified based on comparisons between the learned policy of the agent and policy aggregated from human feedback.

5.3 Hybrid Approaches

As the name suggests, hybrid approaches combine top-down and bottom-up technologies to form a new approach, generally aimed at combining the benefits of both. Inductive logic programming (Cropper & Dumančić, 2022) appears to be the most prominent category of work within hybrid contributions.

5.3.1 Inductive Logic Programming

Inductive logic programming is a technique that aims to generalise a hypothesis represented as a set of logical rules from training examples using machine learning (Cropper & Dumančić, 2022). There are a few works (Anderson et al., 2006; Dyoub et al., 2019; Awad et al., 2020) which utilise inductive logic programming to aid the ethical decision-making process. The ethical advisor, MedEthEx (Anderson et al., 2006), guides healthcare workers who may face ethical dilemmas using Beauchamp and Childress' Principles of Biomedical Ethics (Beauchamp & Childress, 1979) as a basis. A biomedical ethicist trains the system by

supplying the most ethical course of action for a real-life ethical scenario as well as an estimate of the intensity for each of the duties satisfied or violated by an action for all possible actions. Based on the training, the system hypothesises the relationships between the duties from the Principles of Biomedical Ethics. For new ethical situations, the system's knowledge base will interact with an advisor module to determine the degree of severity for violating each ethical duty in the scenario and determine the most appropriate action. An explanation based on the training cases and the refined hypothesis is provided.

By integrating rule-based logic programming, specifically answer set programming for knowledge representation and reasoning, with inductive logic programming to learn from example cases, Dyoub et al. (2019) is able to combine the advantages of top-down and bottom-up approaches. Their work consists of a transparent and accountable knowledge base and is also capable of generating details that may be missed from explicit encodings of rules, which could be necessary for reasoning about future ethics cases. This represents a common type of hybrid approach in combining the advantages of top-down and bottom-up techniques, as also observed in the work by Madl and Franklin (2015) discussed in Section 5.3.2.

The study by Awad et al. (2020) specifically focuses on getting public opinions on the difficult situations where values come into conflict. This paper uses machine learning to extract principles from opinions gathered from the public to guide the behaviour of autonomous systems in an ethical dilemma. Specifically, factors and conflicting features in a scenario are identified and "dilemma vignettes" are created. For example, in an autonomous vehicle scenario, a factor may be one's relation to the autonomous vehicle, whether they are a passenger or a pedestrian. The conflicting features would be sparing the passengers and sparing the pedestrians. Since there may be scenarios where a decision needs to be made between the two, the dilemma vignette, in this case, would be a decision that spares passengers versus a decision that spares pedestrians. Answers to dilemma vignettes for different situations are gathered from the public via a survey and represented using a case-supported, principle-based approach (CPB). An example of this approach can be found in the Moral Machine website.[14] Principles are then extracted using inductive logic programming techniques where rules are generated for machines to interpret, and policies (human-readable rule-equivalents) are produced for humans to understand.

5.3.2 Others

A hybrid approach, as proposed by Madl and Franklin (2015), is based on a biologically inspired cognitive architecture known as LIDA. This proposal stems from the architecture's design, which aims to mimic human cognitive processes and includes mechanisms necessary for moral decision-making. The system stores rules in its memory that are recalled when relevant situations arise. In more difficult situations where rules fail to describe the ethics, the system takes a bottom-up approach. It "understands" emotions through detecting specific situations to activate particular "feeling nodes" within the system. A high enough activation of the nodes leads to performing certain actions corresponding to the node. For example if the system detects one has fallen, it will activate the feeling of concern leading to the action of calling for help.

14. http://moralmachine.mit.edu

Azad-Majiri (Azad-Manjiri, 2014) tries to learn biomedical ethics and abstract relationships between ethical principles, specifically Beauchamp and Childress' Principles of Biomedical Ethics (Beauchamp & Childress, 1979) and actions using a decision tree. The author applies the C4.5 algorithm, a popular method for constructing decision trees, from training data collected from biomedical ethicists. The extracted rules are then utilised for ethical decision-making, ultimately guiding healthcare workers when they encounter ethical dilemmas. We have categorised this particular approach as a hybrid approach due to its consideration of the Principles of Biomedical Ethics as opposed to approaches which purely focus on examples. Instead of a decision tree, Honarvar and Ghasem-Aghaee (2009) integrates an artificial neural network into a BDI model to classify moral and immoral actions into different ethical levels as its hybrid approach.

Also taking on machine learning as the bottom-up approach in its hybrid system, a recent study by Fox and Rey (2024) presents a representation of human ethical requirements using a type of hybrid machine learning model known as Algebraic Machine Learning (AML). AML represents mathematical models consisting of algebraic representations. Concepts relevant to the real-world for a given domain or situation (e.g., location, weather condition for driving) are selected as inputs into the model which are represented as constants used by the algebra. The model is trained on the AML embeddings (based on the constants) and ultimately used for binary classification of ethics.

Instead of machine learning as the bottom-up component of the hybrid approach, Ramanayake and Nallur (2024)'s elder-care robot implementation adopts case-based reasoning and uses expert knowledge to challenge top-down rules if required when making decisions under certain circumstances. The approach enables Pro-Social Rule Bending (PSRB) where specific top-down rules may be temporarily bent (i.e., intentionally violated) in particular situations.

6. EVALUATION Methods

As novel approaches continue to emerge in CME, questions naturally arise about their efficacy and performance. The current state of CME faces a challenge: there is no standardised metric or universally accepted method for evaluating these approaches. This challenge stems from the diverse range of methodologies employed by researchers in the field, each driven by unique goals when instilling ethical considerations into machines. Additionally, the range of ethical problems addressed is incredibly diverse, further complicating the evaluation process. Notably, many researchers create their own criteria to assess the performance of their approach. Moreover, not all proposed works include an explicit evaluation step, contributing to the varied nature of evaluations. In this section, we delve into existing evaluations and highlight recent efforts aimed at advancing the evaluation of approaches within the field.

Table 3: Overview of examples of CME EVALUATION discussed.

Category	Sub-Category	Papers
Criteria	Ethical Performance (Normative)	(L. A. Dennis, Fisher, & Winfield, 2015) (Hendrycks, Burns, et al., 2021)
	Ethical Performance (Descriptive)	(Loreggia et al., 2022) (Hendrycks, Burns, et al., 2021) (Hendrycks, Mazeika, et al., 2021) (Pan et al., 2023)
	Computational Efficiency	(Loreggia et al., 2018) (Lindner et al., 2020) (Stenseke & Balkenius, 2022) (Stenseke, 2024)
Technique	Empirical Methods	(Bremner et al., 2019) (Winfield et al., 2014) (Vanderelst & Winfield, 2018) (Abel et al., 2016) (Loreggia et al., 2018) (Stenseke & Balkenius, 2022) (Jiang et al., 2021)
	Formal Methods	(L. A. Dennis et al., 2015) (L. A. Dennis et al., 2016) (Bremner et al., 2019) (Lindner et al., 2020)

6.1 Criteria to Evaluate Against

One of the primary concerns for researchers in CME is evaluating the system or agent's ability to make ethical decisions in given situations. This evaluation is typically approached in two main ways: (1) based on normative grounds; or (2) descriptive grounds, and is generally associated with the specific SOURCE of guidance (refer to Section 4 for normative and descriptive ethics) that the system has used for its ethical decision-making (Edmond et al., 2022). Additionally, computational efficiency is sometimes considered in CME studies when conducting evaluations. Some researchers choose to focus on one criterion of evaluation while others address multiple criteria in one or more evaluations. We will now discuss works highlighting some of these criteria and the application of these criteria will be addressed in Section 6.2.

6.1.1 Ethical Performance—Normative Evaluation Criteria

Normative evaluations generally focus on whether an agent's behaviour is consistent with a set of defined ethical guidelines, performing as it "should" in an ethical scenario. This relates closely to the SOURCE of guidance for its ethical decision-making (see Section 4). As an example, L. A. Dennis et al. (2015) and Winfield et al. (2014) implement Asimov's first law of robotics: "A robot may not injure a human being or, through inaction, allow a human

being to come to harm" which L. A. Dennis et al. (2015) replicates and proves the properties within scenarios relating to a human falling into a hole (representing danger). For example, "it is always the case that if a_1 is a selected action and its outcome is predicted to be that the human has fallen in the hole, then all the other actions are also predicted to result in the human in the hole". It is important to recognise that, in some cases, a normative evaluation may differ from how humans actually make these decisions where decisions can be impacted by biases.

In a different form, the *ETHICS* dataset introduced by Hendrycks, Burns, et al. (2021) is designed as a benchmark dataset for ethical learning algorithms. Unlike other datasets, which often simply consist of examples of "good" and "bad" behaviours (see Section 4.4 and 6.1.2), the authors intentionally aim to broadly cover different theories within normative ethics. Well-established ethical theories are represented through contextualised everyday scenarios that explicitly relate to the five ethical perspectives: commonsense moral intuitions, deontology, justice, utilitarianism, and virtue ethics. The dataset avoids ambiguities and only consists of scenarios where the ethical choice is clear and unambiguous.

6.1.2 Ethical Performance—Descriptive Evaluation Criteria

Unlike normative evaluations where a set of laid out ethical guidelines is the basis for its criteria of evaluation, descriptive evaluations often examine whether an agent's behaviour aligns with human behaviour (often less explicitly outlined), accurately reflecting human ethical conclusions in various ethical scenarios. Amongst different groups, cultures, and individuals, such ethical conclusions may also vary. Loreggia et al. (2022) is an example of evaluating ethical performance based on whether it mimics human decision-making, comparing decisions of their system through an Amazon Mechanical Turk study.

With the rise in popularity of Large Language Models, in the recent few years, there has also been a rise in the number of benchmark datasets capturing human moral judgements (Hendrycks, Burns, et al., 2021; Hendrycks, Mazeika, et al., 2021; Pan et al., 2023; Reinecke et al., 2023) to be utilised in evaluating ethical decision-making of large language models. For example, Hendrycks, Burns, et al. (2021) collects moral judgements for different scenarios from English speakers across multiple countries for its dataset. It is interesting to note that their dataset has been included as both a normative and descriptive evaluation criteria in this paper due to their deliberate approach in the design of their scenarios to target different ethical theories whilst gathering judgements from different people for these scenarios. Hendrycks, Mazeika, et al. (2021) introduces an environment suite intended to be used in the evaluation of reinforcement learning systems. It consists of complex ethical scenarios with numerous choices of action for evaluating ethical behaviours in text-based games in relation to human values. Various real-life scenarios ranging from negative situations such as theft and animal cruelty and more positive experiences are incorporated. Section 4.4 provides some additional descriptions of various datasets. They not only act as a source for guiding ethical decision-making in some learning systems but can also act as an evaluation measure in others.

6.1.3 Computational Efficiency

Previous studies by Loreggia et al. (2018); Lindner et al. (2020); Stenseke and Balkenius (2022); Stenseke (2024) have rigorously assessed the computational efficiency of ethical decision-making algorithms and approaches. Lindner et al. (2020); Stenseke and Balkenius (2022); Stenseke (2024) mainly focus on evaluations in regards to computational efficiency. Loreggia et al. (2018) evaluate their work using efficiency alongside examining their accuracy in practice according to certain ethical principles which the authors do not specify (their approach is discussed in Section 6.2.1). These evaluations encompassed aspects including computation time and error rate. The significance of these assessments is particularly evident in time-sensitive ethical decision-making, where the speed and the effectiveness of algorithmic decisions are crucial.

6.2 Techniques Used for Evaluation

When it comes to evaluating methods based on criteria including those previously mentioned (see Section 6.1), most studies in the field evaluate their work using either empirical techniques or formal methods. The choice of evaluation technique depends not only on the criteria being assessed but also on the specific decision-making approach adopted by the research. While some studies have no evaluations, other studies may choose to evaluate their approaches in multiple ways, such as (Bremner et al., 2019), formally verifying their program but also conducting experiments to validate their approach. Table 3 presents an overview of the discussed evaluation techniques in CME. Please note that, while the table provides a representation of diverse evaluation approaches, it does not encompass all works discussed in this survey. This design choice prioritises readability, especially considering the extensive volume of literature within the field with many lacking evaluation procedures (see Section 6.3).

6.2.1 Empirical Methods

Experiments are conducted on self-designed scenarios to evaluate the ethical judgments of agents. Depending on the implementation approach chosen by researchers, experiments may involve running programs, models, simulations, or working with physical robots in diverse and unique ethical scenarios and dilemmas. The results of these experiments are then quantitatively compared to human judgments (for descriptive ethics) or qualitatively discussed in relation to the adopted ethical principles (for normative ethics).

Winfield et al. (2014) conducted experiments on e-puck mobile robots involving one actual robot, proxy human(s), and a virtual hole that is dangerous to fall into as the human(s) and robot navigate around. There were three sets of experimental trials: (1) a baseline test, where the task was simply getting the robot to navigate a safe path, avoiding the hole to ensure its own safety; (2) a trial involving the proxy human, where the robot was expected to interact with the human when necessary to prevent them from reaching the hole; and, (3) an additional proxy human introduced to create an ethical dilemma where the robot had to try to ensure the safety of both proxy humans.

Vanderelst and Winfield (2018) conducted experiments on two NAO humanoid robots where one represented the actual robot and the other pretended to be a human. Simple scenarios involving the robot-human travelling to different locations where some locations are dangerous were set up in the experiment. These scenarios targeted self-preservation,

obedience, and human safety, as well as what happens when the robot needs to make trade-offs between human safety and obedience, which are key points in Asimov's laws. The robot successfully adhered to Asimov's Laws of Robotics (Asimov, 1950) in the scenarios. Instead of focusing on the interaction and consequences of ethical situations in a multi-agent environment, Abel et al. (2016) conducted experiments on the Cake or Death problem and the Burning Room dilemma.

Loreggia et al. (2018) provides an empirical analysis of its CP-net approach, which models both ethical principles and subjective preferences of decision makers. The analysis is divided into two parts: the first part examines the model's computation time and error rate; whilst the second part ensures that preferences are not greatly compromised when adhering to ethical principles. An ethical decision is determined based on whether the computed distance (CPD) between the CP-net of subjective preferences, and the CP-net of ethical principles is less than a certain threshold. The experiment generates 1000 pairs of such CP-nets for comparisons. The CPD is compared to the desired distance to determine the accuracy and examine the actual level of sacrifice needed on an individual's preference to be ethical. Stenseke and Balkenius (2022) focuses on assessing the time efficiency of three different ethical algorithms and conducts a qualitative comparison between them. The algorithms are tested in a simulated environment in ten different situations.

Different from the above examples, some works, generally learning-based, evaluate their model against a set of examples to determine its performance. Jiang et al. (2021) is one such example where authors extensively evaluate their model against a benchmark dataset (Hendrycks, Burns, et al., 2021). Hendrycks, Burns, et al. (2021) comprises text-based ethical situations covering concepts related to commonsense morality, duties, justice, virtues, and well-being. The authors of (Jiang et al., 2021) also compare the accuracy of their learning model with few-shot and zero-shot GPT-3 baselines. The evaluation is conducted quantitatively by computing accuracy scores. Moreover, the model's judgments are compared to human judgments collected from Amazon Turk Mechanical workers.

6.2.2 Formal Methods

Formal methods use mathematical models for analysing and verifying software and hardware systems. It rigorously specifies and verifies the behaviour of what systems should do without constraining the approach undertaken to achieve such behaviour (Woodcock, Larsen, Bicarregui, & Fitzgerald, 2009). The goal of formal methods is to ensure the reliability and correctness of systems in accordance with the rigorous specification. This type of technique has been used to ensure the ethical behaviours of some systems in CME. Specifically, the use of *model checking* (Clarke, Grumberg, & Peled, 1999) is observed within the field used to evaluate high level ethical decision-making of reasoning approaches (often BDI approaches). Model checking is the process of checking that specified logical properties hold in all possible executions of the modelled system.

L. A. Dennis et al. (2015) provides an example of formally verifying the correctness of a robot's consequence engine (Winfield et al., 2014) by model-checking different scenarios. It developed a declarative language for specifying consequence engines in its agent infrastructure layer toolkit (AIL). When specifying a system using the AIL, verification can be performed with the AJPF model checker (L. Dennis, Fisher, Webster, & Bordini, 2012). Similarly, other

approaches (L. A. Dennis et al., 2016; Bremner et al., 2019) also verify their implementations of BDI reasoning using the AJPF model checker. Model checking provides a rigorous way to ensure the correctness of programs at a high level, the effectiveness is dependent on other factors, such as the quality of formal specifications, and the program may still have issues at a lower control level. Different to model checking, some approaches use *formal proofs* to verify their methods or claims. For instance, Lindner et al. (2020) uses proofs to demonstrate the different computational complexities of planning under different ethical principles.

6.2.3 No Evaluation

Within the field of CME, a considerable number of publications fall short in presenting evaluations and results for their work. This limitation may be primarily due to the incomplete implementation of proposed ideas (to the extent that they can be executed and tested in an environment), making a thorough assessment difficult. This trend is especially noticeable in approaches that adopt a top-down decision-making methodology. In instances where formal evaluations are absent, some publications briefly review their approach (which we do not consider as an evaluation method) by offering a qualitative exploration of their concepts. This involves guiding readers through their ideas using real-life ethical scenarios as exemplified in the works by (Hooker & Kim, 2018; Bench-Capon, 2020; Limarga et al., 2020; Grandi et al., 2023).

6.3 Problems with Current CME Evaluations

Our exploration reveals a limited number of conducted evaluations on developed CME approaches, specifically regarding their ethical performance. Additionally, there is a notable lack of standardisation in evaluation processes, particularly concerning metrics and benchmarks.

The ethical problems being evaluated in different works are often incomparable, as some evaluate at a higher level, while others focus on lower-level scenarios to verify whether specific decisions produce expected results. Normative and descriptive evaluation criteria are based on the SOURCE of decision-making for the particular DECISION approach and vary vastly depending on the interpretation and choice of the SOURCE theories. Even though evaluation criteria such as computational efficiency are generally comparable, it does not provide insights into the relative significance of the CME approaches in terms of ethics.

Additionally, evaluations span various domains, shaped by the purpose of the work and the choice of SOURCE theories. Technically, there is a general observation from our explorations that approaches that adhere to a normative SOURCE and are evaluated based on normative grounds tend to more commonly be evaluated using some form of technical formal methods whereas approaches that involve descriptive SOURCEs to decision-making would evaluate outcomes based on data that reflect opinions from people such as the general public or responses from ethicists/experts. The evaluation technique heavily depends on and is generally consistent with the type of SOURCE for decision-making (see Section 4).

Table 3 is our best effort in the categorisation of evaluation approaches. Additional patterns and further sub-categorisation beyond those discussed remain difficult to extract due to the scattered and inconsistent evaluations across existing works. We hope this section prompts more research into standardising evaluation metrics to facilitate meaningful

comparisons and growth in the field. Addressing this issue remains a significant challenge for CME at present.

6.3.1 Recent Efforts in Improving CME Evaluations

Recent contributions have made small strides in addressing these challenges by providing valuable tools and resources, such as (Bjørgen et al., 2018; L. Dennis, 2018; Hendrycks, Burns, et al., 2021; Costantini, 2022; Scheirlinck, Chaput, & Hassas, 2023; Akintunde et al., 2023), to assist CME researchers in enhancing the rigour of CME evaluations and comparability of different approaches. For instance, a recent study (Scheirlinck et al., 2023) identifies the lack of a testing environment for ethical decision-making algorithms and takes steps to facilitate such testing by contributing an open-source ready-to-use environment "Ethical Smart Grid" so that CME researchers are able to focus on the core problem of ethical decision-making and have the ability to make comparisons of different contributions in a common environment. Ethical Smart Grid is a simulator designed for testing reinforcement learning ethical decision-making algorithms in the context of a smart grid where agents represent the inhabitants of the smart grid who interact with the grid by consuming or distributing energy whilst considering the ethics (e.g., equity in the distribution of comforts as a result of energy usage). On the other hand, Akintunde et al. (2023) proposes a framework for testing ethical decision-making in autonomous agents consisting of a test-case generation algorithm, an analyser for test results and a learning algorithm that can learn a model of an ethical theory and make adjustments based on feedback. Additionally, both L. Dennis (2018) and Costantini (2022) provide a logic-based toolkit that can assist in ensuring the trustworthiness and ethics of systems.

In addition to the aforementioned benchmark datasets for large language models (Hendrycks, Burns, et al., 2021; Hendrycks, Mazeika, et al., 2021; Pan et al., 2023), Bjørgen et al. (2018) contributes to building benchmarking standards within CME research and development with their open source repository of example scenarios focused on ethical dilemmas. They recognise the cross-disciplinary nature of the field and created this repository with the intention of not only benchmarking but also for researchers of top-down reasoning systems to learn existing examples in the literature. Similarly, Dwivedi et al. (2023)'s work provides a corpus of social norms for analysing the performance of large language models.

Many of the above resources have yet to become popular in the evaluation of CME systems. However, as more such resources have become available in recent years, an increased usage of existing benchmark datasets are observed, fueled by the growing attention towards governing ethics in AI systems and learning models. Like aforementioned, Jiang et al. (2021) evaluates their model against (Hendrycks, Burns, et al., 2021). More generally, a recent study of a comprehensive evaluation of ChatGPT performance (Laskar et al., 2023) utilises (Hendrycks, Burns, et al., 2021) as part of evaluating the ethical performance of ChatGPT.

6.4 Evaluation Across the Source-Decision-Evaluation Taxonomy

Given the challenges surrounding the evaluation of approaches in CME, we contribute to the efforts in evaluations within the field more generally in this section. Specifically, we conducted an evaluation across our Source-Decision-Evaluation taxonomy. This evaluation provides insights into some of the patterns and tendencies, revealing the alignment of certain

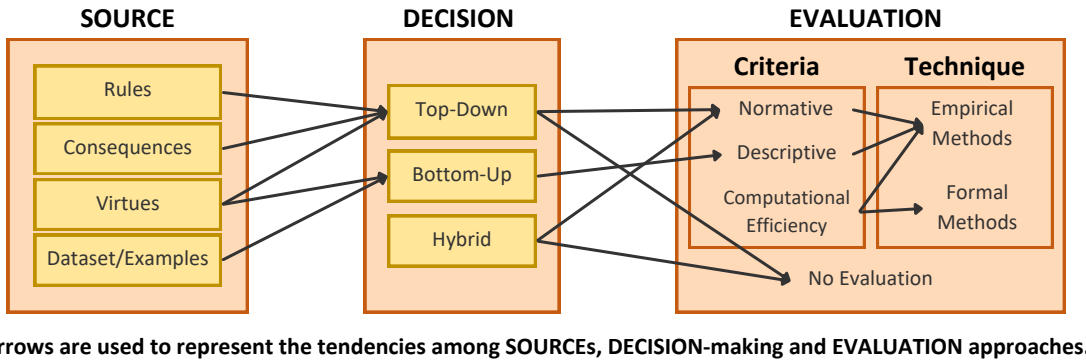

Figure 4: Tendencies among SOURCEs, DECISION-making and EVALUATION approaches.

DECISION-making methodologies with specific SOURCEs and EVALUATION methods. While this is not an evaluation of CME approaches, we believe it offers valuable insights into the trends in the field as a whole and contributes to a broader understanding of where evaluation practices may be lacking within CME approaches. We collated all the papers from Table 1, Table 2, and Table 3 into a single table. Each row represents a paper, with columns detailing the source, decision approach, evaluation criterion, and evaluation technique. Papers spanning multiple categories were accounted for to ensure a comprehensive analysis. Figure 4 illustrates key findings, highlighting the most prevalent tendencies (indicated by the arrows) observed from the table. These patterns are based exclusively on the CME papers included in this survey.

Among papers utilising normative sources (rules, consequences, or virtues), the majority adopt a top-down decision-making approach. Approximately 70% of contributions using rule-based or consequence-based sources employ a top-down methodology. Virtue-based sources, however, are observed to be equally combined with bottom-up methods as with top-down approaches. Given their abstract nature and openness to interpretation, some CME researchers have explored incorporating virtue-based sources through bottom-up methodologies, although the adoption of virtues in general remains rare. The majority of studies (approximately 80%) utilising datasets or examples favour a bottom-up approach.

The choice of sources and decision-making approaches can influence the evaluation criteria and techniques. Top-down approaches, which typically rely on normative sources, naturally align with evaluations based on normative criteria (when an evaluation is conducted), typically carried out empirically. However, nearly 50% of the papers with top-down approaches that we examined lack any form of evaluation. In contrast, most bottom-up approaches include evaluations. These papers often use datasets and examples that reflect human ethical beliefs, so the evaluations are generally based on descriptive criteria, which are frequently assessed empirically. Computational efficiency, a general criterion that is both empirically and formally assessed, is predominantly applied to top-down approaches. However, assessing computational efficiency remains relatively uncommon in CME works.

7. Conclusion

CME has emerged over the past two decades and is advancing at a rapidly accelerating pace. Particularly in recent years, the profound advancement and widespread adoption of AI technology across practical domains have catapulted AI and ethics into the forefront of public and media attention. CME, therefore, stands as an inevitable and critically significant research area. This survey examines the current landscape of CME, highlighting its expansive and inherently multi-disciplinary nature. The review explores the field's diverse methodologies and theoretical foundations that inform ethical decision-making and the evaluation processes. Our systematic organisation of the literature into a *Source-Decision-Evaluation taxonomy* serves to highlight the essential aspects of CME in a modular manner, emphasising the imperative to investigate these facets both independently and collectively. Ongoing advancements are being made by researchers in various facets of CME. Nevertheless, the field is not without limitations, prompting the identification of open challenges and suggestions for future research directions.

7.1 Open Challenges

High-level challenges were identified by McLaren (2006) in the early days of CME. Many of these challenges still persist with questions left unanswered as researchers continue to explore the field. These include:

- The difficulty in connecting abstract ethical rules with specific ethical scenarios;
- There is no universal agreement on the best ethical theory/approach;
- The difficulty in enabling computation without relying on simplifying assumptions or subjective interpretation; and,
- The human ethical decision process is currently not fully understood and seems to be influenced by many variables such as people's culture and beliefs.

As we remain mindful of these overarching challenges, we proceed to examine more specific issues that have emerged in the landscape of CME in recent years as researchers continue to attempt to answer the above questions. We will categorise these challenges using the same framework as our Source-Decision-Evaluation taxonomy, underscoring the necessity for independent investigations within each facet of our taxonomy to prevent potential limitations in our approaches and thought process.

7.1.1 SOURCE

Interdisciplinary Effort Most current approaches in the field are not domain-specific. However, as with any tasks, divide and conquer and limiting the scope may simplify the task of implementing ethics into machines and in turn accelerate the research in this field in various domains. In order to do so, we need to begin from the SOURCE, that is principles and datasets guiding ethical judgments in machines need to be formulated for each domain in society. This would encourage more domain-specific research to be pursued across a greater variety of domains. As observed in CME, there are more works in the medical domain than most other domains and this may be largely motivated by having more established principles

in the domain such as the Principles of Biomedical Ethics (Beauchamp & Childress, 1979). To establish similar standards in other domains, we require collaboration with ethicists, each domain's experts as well as computer scientists. Additionally, as many government and large organisations are beginning to establish ethical principles relating to AI technologies (See Section 4.1.1), their involvement with such collaborations become important for the advancement of society as a whole.

Ethical Principles vs. Social Norms vs. Legal Rules Currently, there exists a grey area in differentiating between ethical principles, social norms, and legal rules within CME approaches. While studies, such as (Loreggia et al., 2018), acknowledge and address subjective personal preferences separately from ethical principles, the explicit distinctions between ethical principles, social norms, and legal rules are often overlooked. Some works even merge the concepts of ethical principles and social norms. For instance, considering the distinction between an ethical principle (e.g., "do no harm") and a social norm (e.g., "shake hands when meeting someone") (Malle et al., 2017), there should intuitively be a difference reflected in the decision-making process. Moreover, Collenette et al. (2022) propose legal reasoning as a special case of ethical reasoning. Recognising and comprehending the discrete differences and relationships among ethical principles, social norms, and legal rules, as well as their roles in the ethical decision-making process, may offer valuable insights within the field.

Robot Behaviour vs. Human Behaviour Whether ethical judgments in machines are guided by rules, consequences, or examples, these sources of guidance are largely derived from human perspectives on ethical behaviour; whether it is what humans "should" do or how humans actually behave. Asimov's Three Laws of Robotics, although flawed and fictional (Asimov, 1950), stand out as one of the few sets of principles describing how robots should behave. The question of whether machines should emulate human behaviour or adhere to alternative guidelines is an area that warrants further exploration in different domains (Vanderelst & Willems, 2020; Vanderelst, Jorgenson, Ozkes, & Willems, 2023). Beyond simply mimicking humans, how do we envision machines behaving? Investigating this direction will prompt reflection on the appropriateness of current sources guiding ethical decision-making in machines for society.

Corpus Variety Currently, bottom-up approaches rely mostly on English text-based example ethical scenarios and judgments from people of an English speaking background. We need more variety for our corpus in which machines can learn from. Awad et al. (2018) is one of the few examples which support numerous languages other than English. As recognised by other researchers in the field, we need to have more inclusive and diverse datasets involving different languages, countries, culture and background. In addition to this, an interesting direction for further research in the future may be to explore curating and learning from datasets with ethical examples of a different format such as images and videos.

7.1.2 DECISION

Hybrid Approach CME employs diverse approaches, each offering unique advantages. Top-down approaches (L. A. Dennis et al., 2016; Van Dang et al., 2017) prioritise explain-

ability and transparency, while bottom-up approaches (Jiang et al., 2021; Vishwanath et al., 2023) excel in flexibility, particularly in novel scenarios. The choice among factors such as accuracy, consistency, efficiency, explainability, transparency, or flexibility in ethical decision-making may vary based on the specific scenario in which the agent operates. Human decision-making, too, involves trade-offs and the selection of different approaches depending on the context. Exploring a hybrid approach (such as (Ramanayake & Nallur, 2024)) becomes a compelling direction for the field, aiming to identify the most crucial factors for a given scenario and applying an appropriate technique accordingly. Moreover, further research into hybrid approaches in general, effectively leveraging the strengths of various techniques is ideal.

Multi-SOURCE Approach While data-driven approaches can be flexible and accurate to a practical extent. It poses inherent controversies in ethically sensitive contexts, especially considering the inherently controversial nature of ethics itself exemplified by controversies and criticisms that surrounded (Jiang et al., 2021) in the media (see Section 5.2.2). The introduction of additional controversies may create complications and pose an unaffordable risk in ethical decision-making for society. Recognising the challenges in the flexibility and practicality of alternative approaches, we assert that decisions based on data should ideally serve as supplementary or a final recourse, particularly in sensitive ethical contexts. In response to this challenge, we propose exploring a hybrid multi-SOURCE approach, where data supplements explicit fundamental ethical principles. The dataset can be curated based on the same theory as the explicit rules, mirroring the developmental process of a young child who learns fundamental values from their parents before gaining insights from personal experiences (drawing an analogy from (Allen et al., 2005)).

Collaboration Difficulties As an evolving domain, CME is currently still in the early stages of exploration, witnessing significant growth as researchers employ a diverse array of techniques to address core challenges. We begin to observe some positive collaborations within specific techniques. For example, for data-driven approaches, researchers are beginning to build on work from one another (e.g., (Jiang et al., 2021; Pyatkin et al., 2023)) and use existing datasets to evaluate their implementations (see Section 6.3.1) instead of working within their own "research groups". However, distinct clusters of techniques have formed within various technical disciplines, limiting collaboration beyond these clusters as observed in our survey. The absence of standardised structures across disciplines poses a challenge, potentially impeding the field's progression in terms of depth. While the breadth of CME continues to expand, there is a need for approaches to converge toward a more focused direction. Researchers are encouraged to explore strategies that foster collaboration and contribute to a more cooperative and dynamic advancement of the field.

7.1.3 EVALUATION

Lack of Standards Expanding upon the challenges presented by the broad scope of the field and collaboration complexities, the comparison and evaluation of different techniques in CME pose significant difficulties. Again, within specific technical disciplines such as data-driven approaches, we are beginning to see a rise in benchmark datasets (Hendrycks, Burns, et al., 2021; Hendrycks, Mazeika, et al., 2021; Pan et al., 2023; Reinecke et al., 2023) in the evaluation of systems. However, we need to extend and establish such standards across the

breadth of CME. Addressing this issue necessitates dedicated research aimed at standardising protocols and establishing metrics crucial to machine ethical decision-making within CME. This imperative task calls for collaborative efforts not only from CME researchers but also from ethicists, domain experts, governmental bodies, and large AI organizations. By collectively identifying essential evaluation criteria, this inclusive approach aims to channel research efforts into a more focused and productive direction.

More Support for Non-Data-Driven Approach Evaluations In Section 6.3.1, we identified some ways in which CME researchers are beginning to support the advancement of evaluation methodologies. However, there is limited support for non-data-driven approaches whether it is in having a technical testing environment or a proposed framework for evaluation. Whether it is in relation to evaluation or other aspects of this taxonomy, there appears to be increasing attention and support on data-driven approaches. Although this aligns with current AI trends, we question whether a solely data-driven approach, although practical, is an appropriate way forward in ethically sensitive contexts. Therefore, it is imperative to sustain efforts in exploring various methodologies.

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society* (pp. 54–61).

Akintunde, M. E., Brandão, M., Jahangirova, G., Menendez, H., Mousavi, M. R., & Zhang, J. (2023). On testing ethical autonomous decision-making. In *Applicable Formal Methods for Safe Industrial Products: Essays Dedicated to Jan Peleska on the Occasion of His 65th Birthday* (pp. 3–15). Springer Nature.

Alexander, L., & Moore, M. (2021). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/`.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*, 149–155.

Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: A comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics*, *2*, 377–387.

Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., & Choi, Y. (2022, July). Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5994–6017).

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, *28*, 15–26.

Anderson, M., & Anderson, S. L. (2008). ETHEL: Toward a principled ethical eldercare system. *AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems*, *Technical Report FS-08-02*, 4–11.

Anderson, M., Anderson, S. L., & Armen, C. (2005). Towards machine ethics: Implementing two action-based ethical theories. In *Machine ethics* (pp. 1–7). AAAI Press.

Anderson, M., Anderson, S. L., & Armen, C. (2006). Medethex: A prototype medical ethics advisor. In *Proceedings of the 18th Conference on Innovative Applications of Artificial*

Intelligence (Vol. 2, pp. 1759–1765).

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI)* (pp. 121–128).

Ashley, K., & McLaren, B. (1994). A CBR knowledge representation for practical ethics. In *European Workshop on Advances in Case-Based Reasoning* (Vol. 984, pp. 180–197).

Asimov, I. (1950). *I, robot.* Gnome Press.

Awad, E., Anderson, M., Anderson, S. L., & Liao, B. (2020). An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence*, *287*, 103349.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*, 59–64.

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., ... Kleiman-Weiner, M. (2022). When is it acceptable to break the rules? Knowledge representation of moral judgement based on empirical data. *Computing Research Repository (CoRR)*, *abs/2201.07763*.

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, *2*, 405–429.

Azad-Manjiri, M. (2014). A new architecture for making moral agents based on C4.5 decision tree algorithm. *International Journal of Information Technology and Computer Science*, *6*, 50–57.

Beauchamp, T. L., & Childress, J. F. (1979). *Principles of Biomedical Ethics.* Oxford: Oxford University Press.

Bench-Capon, T. J. M. (2020). Ethical approaches and autonomous systems. *Artificial Intelligence*, *281*, 103239.

Bendel, O. (2016). Annotated decision trees for simple moral machines. In *Proceedings of the 2016 AAAI Spring Symposium* (pp. 195–201).

Berberich, N., & Diepold, K. (2018). The virtuous machine—Old ethics for new technology? *ArXiv*. Retrieved from `https://doi.org/10.48550/arXiv.1806.10322`

Berreby, F., Bourgne, G., & Ganascia, J.-G. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for Programming, Artificial Intelligence, and Reasoning* (pp. 532–548).

Berreby, F., Bourgne, G., & Ganascia, J.-G. (2017). A declarative modular framework for representing and applying ethical principles. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems* (pp. 96–104).

Bjørgen, E. P., Madsen, S., Bjørknes, T. S., Heimsæter, F. V., Håvik, R., Linderud, M., ... Slavkovik, M. (2018). Cake, death, and trolleys: Dilemmas as benchmarks of ethical decision-making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 23–29).

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics and Information Technology*, *20*, 41–58.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press.

Bourgne, G., Sarmiento, C., & Ganascia, J.-G. (2021). ACE modular framework for computational ethics: Dealing with multiple actions, concurrency and omission. In

International Workshop on Computational Machine Ethics.

Boutilier, C., Brafman, R., Domshlak, C., Hoos, H., & Poole, D. (2004). CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research (JAIR)*, *21*, 135–191.

Bremner, P., Dennis, L. A., Fisher, M., & Winfield, A. F. T. (2019). On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, *107*, 541–561.

Bringsjord, S., & Taylor, J. (2012). The divine-command approach to robot ethics. In *Robot ethics: The ethical and social implications of robotics* (p. 85-108).

Carlucci, F. M., Nardi, L., Iocchi, L., & Nardi, D. (2015). Explicit representation of social norms for social robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4191–4196).

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*, 501–532.

Cheng, L., Mosallanezhad, A., Sheth, P., & Liu, H. (2021). Causal Learning for Socially Responsible AI. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence* (pp. 4374–4381).

Clarke, E., Grumberg, O., & Peled, D. (1999). *Model Checking*. The MIT Press.

Cloos, C. (2005). The Utilibot project: An autonomous mobile robot based on utilitarianism. In *AAAI Fall Symposium—Technical Report.*

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *AAMAS'16: Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1106–1114).

Collenette, J., Dennis, L. A., & Fisher, M. (2022). Advising Autonomous Cars about the Rules of the Road. *Electronic Proceedings in Theoretical Computer Science*, *371*, 62–76.

Costantini, S. (2022). Ensuring trustworthy and ethical behaviour in intelligent logical agents. *Journal of Logic and Computation*, *32*, 443–478.

Cranefield, S., & Dignum, F. (2020). Incorporating social practices in BDI agent systems. In *Engineering Multi-Agent Systems* (pp. 109–126).

Cranefield, S., Winikoff, M., Dignum, V., & Dignum, F. (2017). No pizza for you: Value-based plan selection in BDI agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 178–184).

Cropper, A., & Dumančić, S. (2022, June). Inductive logic programming at 30: A new introduction. *Journal of Artificial Intelligence Research*, *74*, 765–850.

Denis, L. (2007). Kant's formula of the end in itself: Some recent debates. *Philosophy Compass*, *2*, 244–257.

Dennis, L. (2018). The MCAPL framework including the agent infrastructure layer and agent Java Pathfinder. *Journal of Open Source Software*, *3*(24), 617.

Dennis, L., Fisher, M., Webster, M., & Bordini, R. (2012, March). Model checking agent programming languages. *Automated Software Engineering*, *19*, 5–63.

Dennis, L. A., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, *77*, 1–14.

Dennis, L. A., Fisher, M., & Winfield, A. F. T. (2015). Towards verifiably ethical robot behaviour. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence.*

Dimmock, M., & Fisher, A. (2017). *Ethics for A-Level.* Open Book Publishers.

Driver, J. (2006). *Ethics: The Fundamentals.* Malden, MA: Wiley-Blackwell.

Dwivedi, A., Lavania, P., & Modi, A. (2023). EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6921–6931).

Dyoub, A., Costantini, S., & Lisi, F. (2019). Towards ethical machines via logic programming. *Electronic Proceedings in Theoretical Computer Science*, *306*, 333–339.

Dyrkolbotn, S., Pedersen, T., & Slavkovik, M. (2017). Classifying the autonomy and morality of artificial agents. In *First workshop, CARe-MAS@PRIMA 2017* (pp. 67–83).

Edmond, A., Levine, S., Anderson, M., Susan Leigh, A., Conitzer, V., Crockett, M. J., ... Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences*, *26*, 388–405.

Emelin, D., Bras, R. L., Hwang, J. D., Forbes, M., & Choi, Y. (2021). Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 698–718).

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., & Choi, Y. (2020). Social Chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 653–670).

Fox, S., & Rey, V. F. (2024). Representing human ethical requirements in hybrid machine learning models: Technical opportunities and fundamental challenges. *Machine Learning and Knowledge Extraction*, *6*(1), 580–592.

Fraser, K. C., Kiritchenko, S., & Balkir, E. (2022). Does moral code have a moral code? Probing Delphi's moral philosophy. *ArXiv*. Retrieved from `https://arxiv.org/abs/2205.12771`

Ganascia, J.-G. (2007a). Ethical system formalization using non-monotonic logics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 29).

Ganascia, J.-G. (2007b). Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, *9*, 39–47.

Ganascia, J.-G. (2015). Non-monotonic resolution of conflicts for ethical reasoning. In *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations* (pp. 101–118). Springer International Publishing.

Gert, B., & Gert, J. (2020). The definition of morality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/entries/morality-definition/`.

Govindarajulu, N., & Bringsjord, S. (2017). On Automating the Doctrine of Double Effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4722–4730).

Govindarajulu, N., Bringsjord, S., Ghosh, R., & Peveler, M. (2019). Beyond the doctrine of double effect: A formal model of true self-sacrifice. In *Robotics and Well-Being* (Vol. 95, pp. 39–54). Springer.

Govindarajulu, N., Bringsjord, S., Ghosh, R., & Sarathy, V. (2019). Toward the engineering of virtuous machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI,*

Ethics, and Society (pp. 29–35).

Grandi, U., Lorini, E., Parker, T., & Alami, R. (2023). Logic-based ethical planning. In *AIxIA 2022—Advances in Artificial Intelligence* (pp. 198–211).

Guarini, M. (2006, July). Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, *21*(4), 22–28.

Han, T. A., Saptawijaya, A., & Moniz Pereira, L. (2012). Moral reasoning under uncertainty. In *Logic for Programming, Artificial Intelligence, and Reasoning* (pp. 212–227).

Hardin, G. (1974). Living on a lifeboat. *BioScience*, *24*(10), 561–568.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. *ArXiv*. Retrieved from `https://arxiv.org/abs/2008.02275`

Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., ... Steinhardt, J. (2021). What would Jiminy Cricket do? Towards agents that behave morally. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).*

Hindocha, S., & Badea, C. (2022). Moral exemplars for the virtuous machine: The clinician's role in ethical artificial intelligence for healthcare. *AI and Ethics*, *2*, 167–175.

Honarvar, A. R., & Ghasem-Aghaee, N. (2009). An artificial neural network approach for creating an ethical artificial agent. In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)* (p. 290-295).

Hooker, J. N., & Kim, T. W. N. (2018). Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 130–136).

Horty, J. (2001). *Agency and Deontic Logic*. New York: Oxford University Press.

Howard, D., & Muntean, I. (2017). Artificial moral cognition: Moral functionalism and autonomous moral agency. In *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics* (pp. 121–159). Springer.

Jentzsch, S., Schramowski, P., Rothkopf, C., & Kersting, K. (2019). Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 37–44).

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... Gao, W. (2024). AI alignment: A comprehensive survey. *ArXiv*. Retrieved from `https://arxiv.org/abs/2310.19852`

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., ... Choi, Y. (2021). Delphi: Towards machine ethics and norms. *ArXiv*. Retrieved from `https://arxiv.org/abs/2110.07574`

Kolodner, J. L. (1993). *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann Publishers.

Kraut, R. (2022). Aristotle's ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics/`.

Krening, S. (2023). Q-learning as a model of utilitarianism in a human–machine team. *Neural Computing & Applications*, *35*, 16853–16864.

Langman, S., Capicotto, N., Maddahi, Y., & Zareinia, K. (2021). Roboethics principles and policies in Europe and North America. *SN Applied Sciences*, *3*, 857.

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *ArXiv*. Retrieved from `https://arxiv.org/abs/2305.18486`

Li, H., Milani, S., Krishnamoorthy, V., Lewis, M., & Sycara, K. (2019). Perceptions of domestic robots' normative behavior across cultures. In *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 345–351).

Limarga, R., Pagnucco, M., Song, Y., & Nayak, A. (2020). Non-monotonic reasoning for machine ethics with situation calculus. In *AI 2020: Advances in Artificial Intelligence* (pp. 203–215).

Limarga, R., Song, Y., Pagnucco, M., & Rajaratnam, D. (2024). Epistemic reasoning in computational machine ethics. In *AI 2023: Advances in Artificial Intelligence* (pp. 82–94).

Lindner, F., & Bentzen, M. (2018). A formalization of Kant's second formulation of the categorical imperative. *ArXiv*. Retrieved from `https://arxiv.org/abs/1801.03160`

Lindner, F., Bentzen, M. M., & Nebel, B. (2017). The HERA approach to morally competent robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6991–6997).

Lindner, F., Mattmüller, R., & Nebel, B. (2020). Evaluation of the moral permissibility of action plans. *Artificial Intelligence*, *287*, 103350.

Liu, Y. (2022). Consensus-determinacy space and moral components for ethical dilemmas. In *Consensus-determinacy space and moral components for ethical dilemmas* (pp. 1038–1057).

Loreggia, A., Mattei, N., Rahgooy, T., Rossi, F., Srivastava, B., & Venable, K. B. (2022). Making human-like moral decisions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 447–454).

Loreggia, A., Mattei, N., Rossi, F., & Venable, K. B. (2018). Preferences and ethical principles in decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 222–222).

Lourie, N., Bras, R. L., & Choi, Y. (2021). SCRUPLES: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 13470–13479).

Madl, T., & Franklin, S. (2015). Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. In *A Construction Manual for Robots' Ethical Systems* (pp. 137–153).

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In *A World with Robots* (Vol. 84, pp. 3–17). Springer.

McLaren, B. (2003). Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence*, *150*, 145–181.

McLaren, B. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, *21*(4), 29–37. doi: 10.1109/MIS.2006.67

McLaren, B., & Ashley, K. (1999). Case representation, acquisition, and retrieval in SIROCCO. In *Case-based Reasoning Research and Development: ICCBR 1999* (Vol. 1650, pp. 248–262).

Moor, J. (2006, August). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*, 18–21.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., ... Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 839–849).

Nahian, M. S. A., Frazier, S., Riedl, M., & Harrison, B. (2020). Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 124–130).

Nallur, V. (2020). Landscape of machine implemented ethics. *Science and Engineering Ethics*, *26*(5), 2381–2399.

Neto, B. F. D. S., Silva, V. T. D., & Lucena, C. J. P. D. (2010). Using Jason to develop normative agents. In *Proceedings of the 20th Brazilian Conference on Advances in Artificial Intelligence* (pp. 143–152).

Neto, B. F. D. S., Silva, V. T. D., & Lucena, C. J. P. D. (2011). NBDI: An architecture for goal-oriented normative agents. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (pp. 116–125).

Nguyen, T. D., Lyall, G., Tran, A., Shin, M., Carroll, N. G., Klein, C., & Xie, L. (2022). Mapping topics in 100,000 real-life moral dilemmas. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 16, pp. 699–710).

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence* (pp. 1587–1594).

Otterbacher, J., & Manolopoulos, Y. (2023). Machine ethics research: Promises and potential pitfalls. *IEEE Intelligent Systems*, *38*(4), 62–68.

Pagnucco, M., Rajaratnam, D., Limarga, R., Nayak, A., & Song, Y. (2021). Epistemic reasoning for machine ethics with situation calculus. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 814–821).

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., ... Hendrycks, D. (2023). Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 26837–26867).

Pereira, L. M., & Lopes, G. (2007). Prospective logic agents. *Progress in Artificial Intelligence*, 73–86.

Pereira, L. M., & Saptawijaya, A. (2009). Modelling morality with prospective logic. *International Journal of Reasoning-based Intelligent Systems*, *1*, 209–221.

Poszler, F., Portmann, E., & Lütge, C. (2024). Formalizing ethical principles within AI systems: experts' opinions on why (not) and how to do it. *AI and Ethics*, 1–29.

Pyatkin, V., Hwang, J. D., Srikumar, V., Lu, X., Jiang, L., Choi, Y., & Bhagavatula, C. (2023). ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11253–11271).

Quinn, P. L. (1978). *Divine Commands and Moral Requirements*. Oxford University Press.

Rafael H. Bordini, M. W., Jomi Fred Hübner. (2007). *Programming Multi-agent Systems in*

AgentSpeak Using Jason. Wiley.

Ramanayake, R., & Nallur, V. (2024). Implementing pro-social rule bending in an elder-care robot environment. In *Social Robotics* (pp. 230–239). Springer Nature.

Rao, A. S., & Georgeff, M. (1995). BDI agents: From theory to practice. In *Proceedings of the 1st International Conference on Multi-Agent Systems* (pp. 312–319).

Reed, G., Petty, M., Jones, N., Morris, A., Ballenger, J., & Delugach, H. (2016). A principles-based model of ethical considerations in military decision making. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *13*, 195-211.

Reinecke, M. G., Mao, Y., Kunesch, M., Duéñez-Guzmán, E. A., Haas, J., & Leibo, J. Z. (2023). The puzzle of evaluating moral cognition in artificial agents. *Cognitive Science*, *47*(8), e13315.

Rodriguez-Soto, M., Rodriguez-Aguilar, J. A., & Lopez-Sanchez, M. (2022). Building multi-agent environments with theoretical guarantees on the learning of ethical policies. In *Adaptive and Learning Agents Workshop (AAMAS 2022)*.

Roselló-Marín, E., López-Sánchez, M., Rodríguez, I., Rodriguez-Soto, M., & Rodríguez-Aguilar, J. A. (2022). An ethical conversational agent to respectfully conduct in-game surveys. In *Artificial Intelligence Research and Development* (pp. 335–344). Amsterdam: IOS Press.

Rossi, F., & Mattei, N. (2019). Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 9785–9789).

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5477–5490).

Sarker, I. H. (2021, March). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Scheirlinck, C., Chaput, R., & Hassas, S. (2023). Ethical Smart Grid: A Gym environment for learning ethical behaviours. *Journal of Open Source Software*, *8*, 5410.

Singh, L. (2022). Automated Kantian ethics: A faithful implementation. In *KI 2022: Advances in Artificial Intelligence* (pp. 187–208). Springer.

Sinnott-Armstrong, W. (2022). Consequentialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2022/entries/consequentialism/`.

Skelton, A. (2022). William David Ross. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2022/entries/william-david-ross/`.

Stenseke, J. (2023). Artificial virtuous agents: From theory to machine implementation. *AI & Society*, *38*, 1301–1320.

Stenseke, J. (2024). On the computational complexity of ethics: moral tractability for minds and machines. *Artificial Intelligence Review*, *57*(4), 105.

Stenseke, J., & Balkenius, C. (2022). Assessing the time efficiency of ethical algorithms. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Cognition (AIC 2022)* (Vol. 1613). CEUR-WS.

Surendran, V., Melo Cruz, A., Wagner, A., Borenstein, J., Arkin, R., & Chen, S. (2022). Informing a robot ethics architecture through folk and expert morality. In *Proceedings of the 7th International Conference on Robot Ethics and Standards.*

Svegliato, J., Nashed, S. B., & Zilberstein, S. (2021). Ethically compliant sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 11657–11665).

Talat, Z., Blix, H., Valvoda, J., Ganesh, M. I., Cotterell, R., & Williams, A. (2022). On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 769–779). Seattle, United States: Association for Computational Linguistics.

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2017). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, *18*, 1429–1439.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys*, *53*, 1–38.

Torras, C. (2024). Ethics of social robotics: Individual and societal concerns and opportunities. *The Annual Review of Control, Robotics, and Autonomous Systems*, *7*, 1–18.

Tufis, M., & Ganascia, J.-G. (2015). Grafting norms onto the BDI agent model. In *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations* (pp. 119–133). Springer.

Van Dang, C., Tran, T. T., Gil, K.-J., Shin, Y.-B., Choi, J.-W., Park, G.-S., & Kim, J.-W. (2017). Application of Soar cognitive agent based on utilitarian ethics theory for home service robots. In *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)* (pp. 155–158).

Vanderelst, D., Jorgenson, C., Ozkes, A. I., & Willems, J. (2023). Are robots to be created in our own image? Testing the ethical equivalence of robots and humans. *International Journal of Social Robotics*, *15*(1), 85–99.

Vanderelst, D., & Willems, J. (2020). Can we agree on what robots should be allowed to do? An exercise in rule selection for ethical care robots. *International Journal of Social Robotics*, *12*(5), 1093–1102.

Vanderelst, D., & Winfield, A. F. T. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, *48*, 56–66.

Vishwanath, A., Bøhn, E. D., Granmo, O.-C., Maree, C., & Omlin, C. (2023). Towards artificial virtuous agents: Games, dilemmas and machine learning. *AI and Ethics*, *3*, 663–672.

Wiegel, V., & van den Berg, J. (2009). Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, *1*, 233–242.

Wilson, S. (2019). *Natural language processing for personal values and human activities* (Ph.D. thesis). Computer Science and Engineering, University of Michigan.

Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In *Advances in Autonomous Robotics Systems* (pp. 85–96).

Winfield, A. F. T., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE*, *107*, 509–517.

Woodcock, J., Larsen, P., Bicarregui, J., & Fitzgerald, J. (2009). Formal methods: Practice and experience. *ACM Computing Surveys*, *41*, 1–36.

Wu, Y.-H., & Lin, S.-D. (2018). A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence* (pp. 1687–1694).

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)* (pp. 5527–5533).

Zagzebski, L. (2010). Exemplarist Virtue Theory. *Metaphilosophy*, *41*, 41–57.

Zhu, L., Xu, X., Lu, Q., Governatori, G., & Whittle, J. (2022). AI and ethics—Operationalizing responsible AI. In *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership* (pp. 15–33). Springer.