Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions

JESSICA WOODGATE and NIRAV AJMERI, University of Bristol, United Kingdom

Responsible AI must be able to make or support decisions that consider human values and can be justified by human morals. Accommodating values and morals in responsible decision making is supported by adopting a perspective of *macro ethics*, which views ethics through a holistic lens incorporating social context. Normative ethical principles inferred from philosophy can be used to methodically reason about ethics and make ethical judgements in specific contexts. Operationalising normative ethical principles thus promotes responsible reasoning under the perspective of macro ethics. We survey AI and computer science literature and develop a taxonomy of 21 normative ethical principles which can be operationalised in AI. We describe how each principle has previously been operationalised, highlighting key themes that AI practitioners seeking to implement ethical principles should be aware of. We envision that this taxonomy will facilitate the development of methodologies to incorporate normative ethical principles in reasoning capacities of responsible AI systems.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Philosophical/theoretical foundations of artificial intelligence**;

1 INTRODUCTION

The rapid development of AI systems entails the importance of understanding their ethical impact from a human perspective, including notions of responsibility. Responsibility concerns human-centred decisions which take into account social and ethical issues (Dastani and Yazdanpanah [33]). To ensure that AI systems behave in responsible ways, reasoning capacities should support ethical evaluation. Ethical evaluation ought to be a reflective development process incorporating social contexts (Heilinger [66], Weinberg [136]). These considerations are captured under the perspective of sociotechnical systems (STS), which integrate the human element in ethical reasoning (Murukannaiah and Singh [99]).

Within the concept of STS, Chopra and Singh [29] argue that adopting a perspective of *macro ethics* is important. Macro ethics takes a holistic viewpoint, considering social context which includes values (what is important to us in life, Schwartz [115]), norms (standards of expected behaviour, Morris-Martin et al. [96]) and other ethical features. Values are an important aspect of context as they reflect stakeholder preferences (Dubljević et al. [42], Liscio et al. [91]). Norms can be harnessed to help imbue values in systems (Montes and Sierra [95]). Considering context in ethical evaluation should thus include relevant values, norms, and other ethical features (Dignum [39], Soorati et al. [123], Yazdanpanah et al. [139]).

Floridi [52] proposes that ethical evaluation can be understood in terms of *hard ethics* and *soft ethics*. Hard ethics is what is morally right in shaping the law by reference to values, norms, rights, duties, and responsibilities. However, there may be cases where ambiguities arise that hard ethics cannot provide an answer for. Stakeholders may have different value preferences, or their values may conflict with norms (Jakesch et al. [74]). Soft ethics examines what ought to be done over and above existing norms, such as in cases where competing values and interests need to be balanced, or existing regulations provide no guidance (Floridi [53]).

Improving the capacity of AI to reason responsibly, considering soft ethics and social contexts, is aided by appeal to normative ethics. Normative ethics is the study of practical means to determine the ethicality of actions through the use of principles and guidelines, or the rational and systematic study of the standards of right and wrong (Murukannaiah and Singh [99]). We argue that operationalising

Authors' address: Jessica Woodgate, jessica.woodgate@bristol.ac.uk; Nirav Ajmeri, nirav.ajmeri@bristol.ac.uk, University of Bristol, Bristol, United Kingdom, BS8 1UB.

rules from normative ethics is a step forward to creating responsible AI which accommodates social contexts in ethical evaluation.

1.1 Motivation for a Taxonomy of Ethical Principles

The motivation for this work stems from the need to improve ethical evaluation capacities for responsible AI. To aid this we look to normative ethics, as engaging with interdisciplinary insights encourages more inclusive and critical thinking (Weinberg [136]). Principles from normative ethics imply certain logical propositions which must be true for a given action plan to be ethical (Kim et al. [81]). Principles can be used to methodically think through dilemmas and promote satisfactory outcomes (Canca [25], Saltz et al. [113]). Operationalising normative ethics principles thereby enables systems to methodically reason about ethics (Woodgate and Ajmeri [138]).

Normative ethical principles have previously been utilised for a variety of AI applications. Binns [16] and Leben [85] apply ethical principles to improve fairness considerations for binary machine learning algorithms. Cointe et al. [30] implement ethical principles in decision making, enabling agents to make ethical judgements in specific contexts. Conitzer et al. [32] state that principles can be applied to identify morally relevant features of dilemmas, whilst Heilinger [66] illustrates how principles frame discussions of risks and opportunities in AI.

Ethical thinking should be fostered through the appreciation of a variety of approaches, considering the strengths and limitations of each (Burton et al. [24]). Adopting interdisciplinary perspectives also helps to bridge epistemic divides (Weinberg [136]). We suggest that a taxonomy of ethical principles previously seen in AI and computer science, including how they have previously been operationalised, provides practitioners with key themes and examples to help ground their approaches. We envision that this taxonomy will contribute to improving ethical evaluation capacities of responsible AI.

1.2 AI Principles and Ethical Principles

In the context of AI and ethics, there are two types of principles referred to: (1) those inferred from normative ethics such as deontology and consequentialism, as found in Leben [85], and (2) those adapted from other disciplines like medicine and bioethics such as those suggested by Cheng et al. [28], Fjeld et al. [50], Floridi and Cowls [54], Jobin et al. [76], Khan et al. [80], and Whittlestone et al. [137], including beneficence, non-maleficence, autonomy, justice, fairness, non-discrimination, transparency, responsibility, accountability, safety and security, explainability, human control of technology, and promotion of human values.

To ensure clarity of terminology, we refer to principles from normative ethics as *ethical principles*, and those highlighted by Floridi and Cowls [54] and Jobin et al. [76] as *AI principles*. We define ethical principles and AI principles as follows:

Ethical Principles. Operationalisable rules inferred from philosophical theories which imply logical propositions denoting moral acceptability.

AI Principles. Ends which ought to be promoted in the development and deployment of AI to ensure it is socially beneficial.

1.2.1 Ethical principles. Ethical principles are philosophical theories which are normative in the sense that they are prescriptive, denoting how things should be, rather than descriptive, denoting how things are (Kim et al. [81]). As what is the case might not be ethical, using independently justified principles has the benefit of addressing the *is-ought* gap: just because something is the case, does not mean that it ought to be. Ethical principles guide normative judgements, determine the moral permissibility of concrete courses of action and help to understand different perspectives

(McLaren [92]). Using ethical principles makes explicit the normative assumptions underlying ethical choices, improving propensity for accountability (Fazelpour et al. [49], Lechterman [86]). Ethical principles can be operationalised in reasoning capacities as they imply certain logical propositions which must be true for a given action plan to be ethical, and provide frameworks for guiding judgement and action (Boddington [18]). The abstractness of ethical principles entails that they can be used to analyse concrete courses of actions in a wide range of situations (Binns [16], Conitzer et al. [32], Lindner et al. [90]).

Ethical principles principles broadly divide into *deontological* principles (those which entail conforming to rules, norms and laws, Hagendorff [62]), *virtue ethics* (denotes moral character central to ethical action, Wallach and Vallor [135]), and *consequentialist* principles (those which derive morality from the outcome of actions, Horta et al. [72]).

1.2.2 AI principles. We understand AI principles as ends which ought to be promoted in the development and use of AI. AI principles are qualities that we should expect AI to embody and by which we can assess how socially beneficial AI is.

1.2.3 Distinction between ethical principles and AI principles. Translating AI principles into practice is challenging (Zhou and Chen [142]). AI principles do not provide guidance for how they can be implemented, and interpretation of their meaning may diverge (Munn [97]). Ethical principles, on the other hand, are abstract rules that provide logical propositions denoting which actions are morally acceptable. Applying ethical principles to indicate moral acceptability helps to determine which actions are aligned with AI principles. Ethical principles are thus abstract rules which can be used to promote the instantiation of AI principles.

To illustrate the distinction, we explore how the ethical principle of egalitarianism helps to implement the AI principle of fairness. Egalitarianism supports the notion that human beings are in some fundamental sense equal (Binns [16]). Fairness is defined by Jobin et al. [76] as the mitigation of unwanted bias and discrimination. To work towards fairness, egalitarianism may be operationalised by reducing inequality to mitigate discrimination. For example, this could take the form of a rule that opportunities must be equally open to all applicants, as seen in Lee et al. [87].

1.3 Gaps in Related Research

Existing taxonomies and surveys are present in the relevant but distinct domain of AI principles, such as Floridi and Cowls [54], Jobin et al. [76], and Khan et al. [80]. However, these works do not consider ethical principles. The rest of this paper therefore surveys ethical principles rather than AI principles. Dignum [40], Leben [85], and Robbins and Wallace [110] provide summaries of normative ethics. Tolmeijer et al. [130] give an overview of implementations of machine ethics, providing useful guidance as to the technical and non-technical aspects of implementing ethics and evaluating systems. Similarly, Yu et al. [141] provide a concise guide to ethical dilemmas in AI and identify a high-level overview of ethical principles. From a philosophical perspective, Boddington [18] presents a comprehensive exploration of the application of three major normative ethics theories (deontology, virtue ethics and consequentialism) to AI, and issues that might arise.

We expand upon previous literature to address gaps concerning principles which were not included in previous reviews, and provide further detail about how ethical principles have been operationalised in AI and computer science.

1.4 Novelty

We build upon previous research, especially Tolmeijer et al.'s [130], to collate a broader range of ethical principles discussed in the AI and computer science literature, summarising operationalisation principle by principle. There are three key aspects of novelty contributed by this paper:

Broadening the Range of Ethical Principles We create a taxonomy tree with 21 ethical principles discussed in AI and computer science literature.

Principle Specific Operationalisation We define a new mapping of each principle to how they have been operationalised in literature. Operationalisation is explained on both an abstract level, including how each principle has been defined in literature and difficulties that may arise, and on a technical level, including technical implementations of each principle, and how technical implementation relates to different architectures.

Reflection on Research Gaps and Directions We identify gaps and future directions. Broadly, directions emerge from (1) expanding the taxonomy to include principles under-utilised in AI and computer science, (2) resolving ethical dilemmas where principles conflict or lead to unintuitive outcomes, and (3) incorporating ethical principles in STS considering broad social contexts.

1.5 Organisation

Section 2 explains our methodology in brief. This will be useful for future research seeking to expand the taxonomy of ethical principles by reproducing the methods used here. Section 3 explores our findings for which ethical principles have been proposed in AI and computer science literature. Section 4 examines how ethical principles have previously been operationalised, and steps practitioners seeking to operationalise principles should take. Section 5 identifies gaps and future directions for operationalising ethical principles in AI and computer science. Section 6 concludes the paper.

2 METHODOLOGY

Taking inspiration from software engineering research, for reproducibility we follow Kitchenham and Charters [82] guidelines on conducting a systematic literature review to develop our taxonomy for ethical principles. We first define our objective and research questions to help scope the search. We construct an initial search string from preliminary research. Using a forwards and backwards snowballing technique, we search selected resources (the University of Bristol library, with Google Scholar as backup) using our search string. We apply inclusion and exclusion criteria to identify primary studies, and follow relevant citations to expand the search. We update the search string if we identify new key words (i.e., if studies reference ethical principles not previously seen), repeating the process until no new key words emerge. For further details of the methodology, see Appendix A.

2.1 Objective

We investigate the current understanding of ethical principles in AI and computer science, and how these principles are operationalised. Specifically, we address the following questions:

Q_p (Principles). *What ethical principles have been so far proposed in AI and computer science literature?*

The purpose of this question is to aid the identification of principles currently used in literature within the domain of AI and computer science. Due to the intricacies of philosophical discourse, we follow Tolmeijer et al.'s [130] approach in providing brief overviews of how each principle has been defined in literature. We do not attempt to give an introduction to moral philosophy, which can be found in works such as Boddington [18].

Q_o (Operationalisation). *How have ethical principles been operationalised in AI and computer science research?*

This question looks at the identified principles to examine how they have been operationalised in AI and computer science. Works such as Leben [85] and Tolmeijer et al. [130] offer guidance as to how specific ethical principles may be operationalised. We expand upon the range of principles presented in previous works.

Q_g (Gaps). *What are existing gaps in ethics research in AI and computer science, specifically in relation to operationalising principles in reasoning capacities?*
This question aids analysis of existing gaps in operationalising the principles in reasoning capacities of responsible AI, to direct future research.

2.2 Relevant Works

We conducted an initial search on 2022-May-23, a second search on 2023-January-14, and a third search on 2024-February-01. The first search produced 3.74 million results on Google Scholar and 998,613 results on the University of Bristol Online Library. Looking at the first 5 pages of results, we applied the inclusion and exclusion criteria, which led to around 10–20 studies from each resource. Closer examination of these works resulted in the identification of relevant citations which we incorporated into our review. The selection of these works was critiqued by a secondary researcher which helped to identify further relevant research. This resulted in 57 papers being included in the review. The second search resulted 10 more papers being included in the review. The third search identified a further 14 papers to include in the review.

3 TAXONOMY OF ETHICAL PRINCIPLES

We now address Q_p (Principles) on identifying ethical principles proposed so far. We first present an overview of principles we identify in AI and computer science literature. We categorise papers based on principles they explicitly mention, their contribution and evaluation type. We then present our findings for each principle, summarising their definition, previous application, and potential difficulties which may arise.

Within normative ethics, there are three main strands of theory: *deontology*, *virtue ethics*, and *consequentialism*. There is a debate as to whether consequentialism and virtue ethics are branches of teleology or distinct branches of theory, as summarised by Spielthenner [124] and further explored by Horta et al. [72]. Following Horta et al. [72] and Boddington [18], we do not use the term teleology, categorising consequentialism and virtue ethics as distinct branches. However, our key intention is to examine how such principles have been used in AI and computer science literature. Further exploring the philosophical relation of these theories is outside the scope of this work.

Deontological theories revolve around rules, rights, and duties (Murukannaiah and Singh [99], Wallach et al. [134]). Virtue ethics denotes that ethicality stems from the inherent character of an individual, not the rightness or wrongness of individual acts (Yu et al. [141]). Consequentialist theories emphasise that whether something is right or wrong depends completely on its outcome (Horta et al. [72]). Figure 1 displays the taxonomy of principles identified in literature in a tree structure, mapping out how they relate to each other.

3.1 Overview of Paper Categorisation

We categorise papers identified in our review based on contributions of the paper, type of evaluation, and ethical principles explicitly mentioned. Expanding on previous work, we adapt Yu et al.'s [141] and Tolmeijer et al.'s [130] taxonomies to categorise papers by contribution and evaluation type. We categorise papers by which principle(s) they explicitly mention with the exception of three papers. Jiang et al. [75] and Shi et al. [119] do not explicitly state ethical principles, but utilise the Hendrycks et al.'s [68] dataset which consists of scenarios based on ethical principles. We include Jiang et al. [75] and Shi et al. [119] as they provide valuable demonstrations for how ethical principles can be implemented. In addition, Noothigattu et al. [102] do not explicitly state ethical principles, but provide a methodology to learn ethical constraints from examples of correct behaviour. We include this paper because it demonstrates how inverse reinforcement learning, a valuable technique for

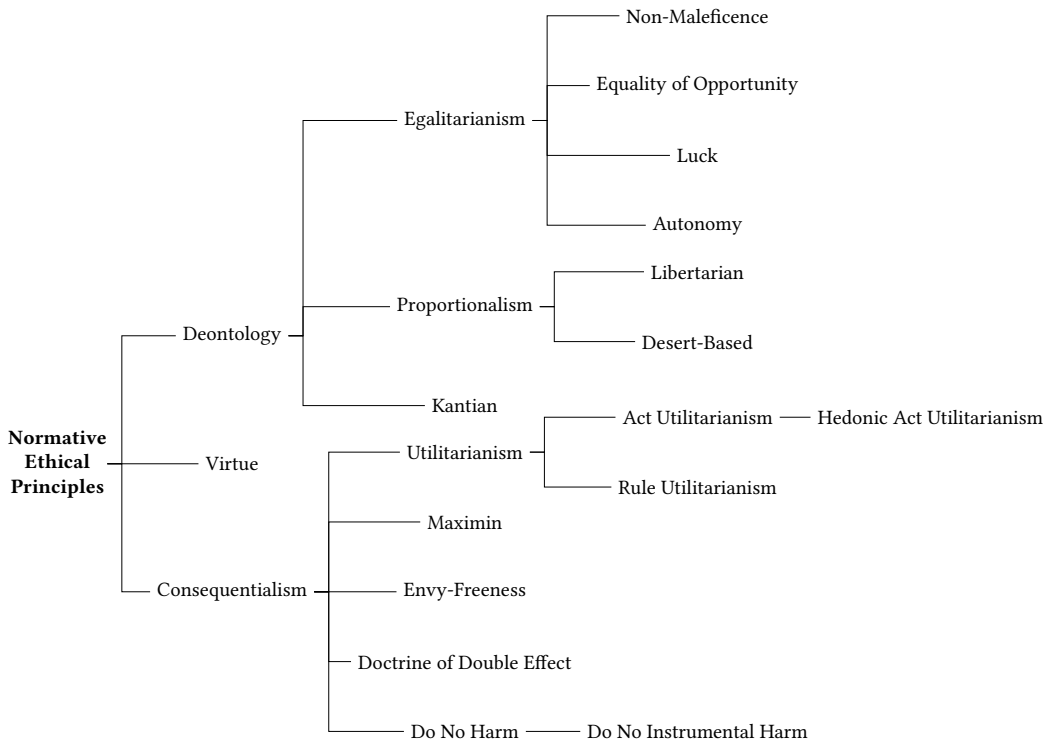

Fig. 1. Taxonomy of ethical principles found in AI and computer science literature

learning ethics through observation, can be used to implement ethical rules (where rules are integral to deontological approaches) in machines. Figure 1 displays the principles we identify.

Based on principles explicitly mentioned, works are broadly categorised into eleven key principles (deontology, egalitarianism, proportionalism, Kantian, virtue, consequentialism, utilitarianism, maximin, envy-freeness, doctrine of double effect, and do no harm).

We categorise six types of contribution: descriptive, model representation, individual ethical decision making frameworks, centralised collective ethical decision making frameworks, decentralised collective ethical decision making frameworks, and ethics in human-AI interaction frameworks. Descriptive papers abstractly evaluate how normative ethics relates to AI. Model representation examines how to appropriately represent ethical knowledge in a system, or what features an ethical system should include. Individual decision making examines how individual agents may judge or select their own actions or the actions of others. Centralised collective decision making involves a central mechanism which makes decisions concerning multiple agents. Decentralised collective decision making involves multiple agents making distributed ethical decisions, such as in multi-agent systems (MAS). Ethics in human-AI interaction investigates ethical considerations of agents designed to influence human behaviours or work in conjunction with humans.

Adapting Tolmeijer et al.'s [130] categorisation, we categorise four evaluation types: test, proof, informal, and none. Test involves empirical analysis, comparing the system outcome against some ground truth. Proof examines if the system behaves according to some known specifications, typically using logic. Informal evaluation compares the system to example scenarios or application domains. When none of these evaluation types pertain, papers are categorised as 'none'.

Tables 1 and 2 display the categorisation of papers. Appendix A.2 provides further details of the methodology for paper classification.

Table 1. Contribution categorisation for deontological principles and virtue ethics. Papers are categorised by their contribution and evaluation type. For contribution, descriptive abstractly explores ethical principles and AI; model representation examines representing ethical knowledge; individual decision making explores individual agents judging or selecting actions; centralised collective decision making involves a centralised mechanism concerning multiple agents; decentralised collective involves multiple agents making distributed ethical decisions; human-AI frameworks investigate agents designed to influence human behaviours or work in conjunction with humans. For evaluation, test involves empirical analysis; proof examines if the system behaves according to some specifications; informal compares the system to example scenarios; none is used when we do not identify an evaluation.

Contribution		Ethical Principles				
Contribution Type	Evaluation Type	Deontology	Egalitarianism	Proportionalism	Kantian	Virtue
Descriptive	None	[1, 14, 16, 24, 62, 66, 79, 113, 125]	[16, 66, 106]	[106]	[1, 14, 26, 65]	[1, 14, 24, 62, 65, 66, 79, 113, 125]
Model Representation	Test	[6, 102]	–	–	–	–
	Proof	[15]	[55, 84]	[84]	[15]	[60]
	Informal	[3, 4, 107]	[9]	[32]	[4]	[3, 63, 107]
	None	[5, 18, 39, 40, 47, 56, 64, 99, 130, 134, 141]	[87, 98]	[47, 87]	[18, 39, 40, 48, 56, 83, 111, 130, 134]	[18, 40, 99, 104, 111, 130, 134, 135, 141]
Individual Decision Making	Test	[35, 70, 75, 112, 119]	–	–	[81, 112, 129]	[70, 75, 112, 119]
	Proof	[89, 90]	–	–	[89]	–
	Informal	[11, 30]	–	–	–	[30]
	None	[141]	–	–	[7]	[141]
Centralised Collective Decision Making	Test	[68]	[34]	–	–	[68]
	Proof	–	[21, 44]	–	–	–
	Informal	[85]	[85]	[85]	–	–
	None	–	–	[108]	–	–
Decentralised Collective Decision Making	Test	[61]	–	–	–	[61]
	Proof	–	–	–	–	–
	Informal	[110]	–	–	[110]	[110]
	None	[141]	–	–	–	[141]
Human-AI Interaction	Test	[6]	–	–	[129]	–
	Proof	–	–	–	–	–
	Informal	[107]	–	–	–	[63, 107]

Contribution		Ethical Principles				
Contribution Type	Evaluation Type	Deontology	Egalitarianism	Proportionalism	Kantian	Virtue
	None	[46, 141]	–	–	[7, 46]	[131, 141]

Table 2. Contribution categorisation for consequentialist principles.

Contribution Type		Ethical Principles					
Contribution Type	Evaluation Type	Consequentialism	Utilitarianism	Maximin	Envy-Freeness	Doctrine of Double Effect	Do No Harm
Descriptive	None	[1, 14, 65, 66, 113, 125]	[1, 24, 79, 113, 118, 125]	–	–	–	–
Model Representation	Test	–	–	–	–	[17]	
	Proof	[15]	[15, 84]	–	–	[59]	–
	Informal	[3, 4, 107]	[3, 4, 134]	[9, 15]	–	[15]	–
	None	[18, 39, 40, 48, 130, 135, 141]	[5, 18, 39, 40, 47, 48, 56, 83, 98, 99, 111, 135, 141]	[87]	–	[40]	[39]
Individual Decision Making	Test	[112]	[2, 35, 70, 75, 81, 112, 119, 129]	[2]	–	–	[37]
	Proof	[89]	[8, 89, 90]	–	–	[90, 94]	[90]
	Informal	[30]	[11]	–	–	–	–
	None	[141]	[7, 141]	–	–	–	–
Centralised Collective Decision Making	Test	–	[13, 68, 88, 117]	[13, 38, 88]	–	–	–
	Proof	–	[27]	[27, 105, 127]	[127]	–	–
	Informal	[85]	[85]	[85]	–	–	–
	None	–	–	–	[19]	–	–
Decentralised Collective Decision Making	Test	[61]	[100]	–	–	–	–
	Proof	–	[58]	[58]	–	–	–
	Informal	–	[110]	–	–	–	–
	None	[141]	[141]	–	–	–	–
Human-AI Interaction	Test	–	[20, 129]	–	–	–	–
	Proof	–	–	–	–	–	–
	Informal	[107]	–	–	–	–	–

Contribution Type		Ethical Principles					
Contribution Type	Evaluation Type	Consequentialism	Utilitarianism	Maximin	Envy-Freeness	Doctrine of Double Effect	Do No Harm
	None	–	[7]	–	–	–	–

We find that certain principles, such as utilitarianism, are more commonly discussed than other principles such as do no harm, as can be seen in Table 2. We also find that there is a large amount of research referencing 'deontology' and 'consequentialism' as broad terms, but not specifying what types of deontology or consequentialism they are referring to, for example, Anderson and Anderson [6], Cointe et al. [30], and Greene et al. [61]. Precisely stating the ethical principles used (e.g., type of deontology) would allow for more exact operationalisation.

In terms of contribution, we find a large majority of works utilise ethical principles in descriptive and model representation papers. Descriptive papers include overviews of ethics, such as Boddington [18], or ethical critiques with reference to ethical principles, such as Heilinger [66]. Model representation harnesses ethical principles to examine which ethical features should be considered in models, such as Anderson and Anderson [5], Dignum [40], and Lee et al. [87]. For individual decision making we find that more works implement consequentialist principles, especially utilitarianism, than deontology or virtue ethics. The majority of these works, for example, Dehghani et al. [35], Svegliato et al. [129], and Ajmeri et al. [2], also provide tests. In centralised collective decision making approaches we find more works implementing consequentialist principles rather than deontology or virtue ethics, such as Bakker et al. [13], Lera-Leri et al. [88], and Patel et al. [105]. For decentralised collective decision making, we find that most works implement deontology, as seen in Greene et al. [61], or utilitarianism, as seen in Governatori et al. [58]. In human-AI interaction we find more works implementing deontological principles or virtue ethics rather than consequentialism, such as Pflanzer et al. [107], Hagendorff [63], and Anderson and Anderson [6]. We find decentralised collective decision making and human-AI interaction involved in the least number of papers, suggesting avenues for future work in these areas.

3.2 Deontology

Deontology entails conforming to rules, laws, and norms, and respecting relevant obligations and permissions that stem from duties and rights (Cointe et al. [30], Hagendorff [62], Rodriguez-Soto et al. [112]). For Deontological theories, the permissibility of action lies within the intrinsic character of the act itself (Boddington [18]). An action is permissible if and only if the act itself is intrinsically morally good, independent of the outcome (Lindner et al. [90]).

To implement deontological theories, a rules-based approach may be used to provide moral orientation in identifying appropriate actions (Heilinger [66]). An example of a rules-based approach is Limarga et al. [89], who use predicates to encode rules and then reason about different types of actions. Berreby et al. [15] implement different deontological specifications as rules in a 'model of the right'. The model of the right is used to generate a 'rightness assessment' of available actions, considering context. The model contains deontological principles in conjunction with consequentialist principles. Similarly, Pflanzer et al. [107] suggest implementing deontology as part of a model which utilises consequentialism and virtue ethics, in which the role of deontology is to analyse the intention of actions. Tolmeijer et al. [130] argue that deontology could be implemented by inputting the action, using rules and duties as the decision criteria, and then mechanising actions via the extent to which they fit with the rule.

Deontology has been applied to different contexts. Binns [16] uses deontology to choose between incompatible fairness metrics, whereas Leben [85] applies it to evaluate distributions of binary classification algorithms. Hendrycks et al. [68] implement deontological principles in contextualised scenarios to measure the ethical knowledge of natural language processing models. Jiang et al. [75] use this dataset to test a model trained on people's moral judgements of various situations. Some works suggest using deontology only in specific circumstances: Dehghani et al. [35] choose to implement deontology in situations with 'sacred values', selecting the action that doesn't violate a sacred value.

However, issues may arise when applying deontology. One common concern is that because deontological approaches focus on the intrinsic nature of an action, they fail to take the most likely consequences into account. Focusing solely on the intrinsic nature of action makes it challenging for deontology to adequately capture complex ethical insights (Abney [1], Saltz et al. [113]). The complexity of ethical insight entails that any system of rules requires some interpretation and understanding of background assumptions. This means that the same rules might be interpreted differently in different contexts or by different people (Boddington [18]). In addition, rights-based ethics revolve around decisions based on the rights of those who are affected by the decision. Focusing on rights is less helpful in situations where rights are not impinged, yet some sort of ethical dilemma is still occurring. For example, spreading hate speech does not necessarily infringe the rights of others, and there are arguments that preventing it infringes the right to free speech. However, there is an intuition that hate speech is wrong. Issues may arise with implementation when exceptions to rules emerge. Rules are expected to be strictly followed, implying that for every exception they must be amended, which could result in very long rules. Determining the right level of detail is thus important to ensure interpretability for the machine (Tolmeijer et al. [130]). Lastly, there may be conflicts between rules. Conflicts may be addressed by ordering or weighing the rules, but this gives rise to difficulties in determining the order of importance.

3.2.1 Egalitarianism. Egalitarianism stems from the notion that human beings are in some fundamental sense equal. To administer egalitarianism, Binns [16] recommends that efforts should be made to avoid and correct certain forms of inequality. Heilinger [66] indicates that egalitarianism can be understood in a relational sense, aiming for conditions under which all can relate to and interact with one another on an equal footing.

Literature implements egalitarianism by promoting equality in different ways: Murukannaiah et al. [98] suggest minimising disparity across stakeholders with respect to satisfying their preferences; Dwork et al. [44] classify individuals who are similar with respect to a particular attribute similarly. For resource allocation, Leben [85] confers equal rights (and thus equal shares) to each member of the population. If achieving equality across all metrics for the entire population is impossible, they suggest a distribution that minimises the distance to some fairness standard.

We identify different applications in which egalitarianism has been implemented. Lee et al. [87] utilise egalitarianism to evaluate various algorithmic fairness metrics, such as predictive parity or equal odds. Applying egalitarianism to fairness metrics helps AI practitioners decide what layers of inequality should (not) influence a model's prediction. Persson and Hedlund [106] suggest utilising egalitarianism to consider how to distribute responsibility for ethical AI development. Botan et al. [21] apply egalitarianism to judgement aggregation.

Certain difficulties are important to acknowledge when considering egalitarianism. Binns [16] highlights a prominent debate as to whether a single egalitarian calculus should be applied across different social contexts, or if there are internal 'spheres of justice' in which different fairness metrics may apply, and between which redistributions might not be appropriate. Particular measures of egalitarianism might apply differently to different contexts. For example, universally enforcing a

literacy test before being allowed to vote for a political election may lead to people from backgrounds with less access to formal education being excluded. However, literacy tests for a job position may seem appropriate if everyone has an equal opportunity to take the test, as talents and abilities vary between individuals. One should thus carefully evaluate the metrics being used to impose egalitarianism. Table 3 describes sub-types of egalitarianism.

Table 3. Sub-types of Egalitarianism.

Principle	Description	Difficulties
Non-Maleficence	Imposes egalitarianism across harms but not benefits (Leben [85]). In optimisation techniques, different actions could be assigned values based on a predetermined formula, identifying harms caused by each action. The action with the most equal distribution of harm is chosen.	Allows for arbitrarily large inequalities in outcomes, and assumes a dubious distinction between 'better-off' and 'worse-off' (Leben [85]). It thus is difficult to define what a harm is and what a benefit is.
Equality of Opportunity	Negative attributes due to an individual's circumstances of birth or random choice should not be held against them. However, individuals should be still held accountable for their actions (Dwork et al. [44], Friedler et al. [55]). Opportunities should therefore be equally distributed. Binns [16] proposes that one could examine whether each group is equally likely to be predicted a desirable outcome, given the base rates for that group. Lee et al. [87] suggests ensuring opportunities are equally open to all applicants based on a relevant definition of merit.	Fleurbaey [51] argues that this can be fully satisfied even if only a minority segment of the population has realistic prospects of accessing the opportunity.
Luck	Inequalities that stem from unchosen aspects should be eliminated so no-one is worse off due to bad luck. Instead, people should receive benefits as a result of their own choice (Dworkin [45], Lee et al. [87]). From an optimisation perspective, people could be given a weighting which mitigates the effects of luck. Allocations are distributed equally, accounting for this weighting.	Defining what is within an individual's genuine control is often difficult (Binns [16]). The ideal solution would allow inequalities resulting from people's free choices and informed risk-taking, disregarding those which are the result of brute luck.
Autonomy	Levels of autonomy should be equally distributed, through a variety and quality of options, and decision-making competence (Fleurbaey [51]). The aim of this would be to incorporate the full range of individual freedom (Lee et al. [87]). Levels of autonomy could be inputted to reason about potential actions, selecting the action with the most equal distribution of autonomy.	When there is a significant asymmetry of power and information, autonomy in rational decision-makers fails as an ethical objective (Fleurbaey [51]).

3.2.2 Proportionalism. Proportionalism entails adjusting the rights of each person proportionally based on their contributions to production. Depending on the sub-type of proportionalism (shown in Table 4), contributions could include the resources from each member of the population that went into production, the amount of actual work that went into the deployment of those resources, or the amount of contribution discounting for luck that went into those resources.

Previous operationalisation of proportionalism includes Leben [85], which constructs utility functions that evaluate the distribution of rights in accordance with contribution. A fairness standard establishes the ideal distribution of rights by dividing the total amount of contribution by each individual's amount of contribution. The best distribution is the one with the minimum distance from this fairness standard for all individuals. Pitt [108] argues that ability to contribute should be central to methodological design of STS; to ensure self-determination, communities must be able to own and operate the platforms they use. Alternatively, Lam et al. [84] assign distance to resource location as proportional to group size.

A challenge with proportionalism is that there may be situations where groups or individuals do not confer contributions to production, but should be granted a distribution of rights. For example, those unable to contribute due to disability should still have a fair distribution of rights. Accommodating those who were unable to contribute may be mitigated by considering the influence of luck. Table 4 shows sub-types of proportionalism.

Table 4. Sub-types of Proportionalism.

Principle	Description	Difficulties
Libertarian	Libertarianism emphasises the importance of each person's freedom (Lee et al. [87]). Rights are distributed according to each person's total contribution at the time of consent. Inequality within the range of this initial contribution is not considered unfair (Leben [85]).	Libertarianism does not target pre-existing inequalities which may be worth mitigating. For example, the contribution of some people may be inhibited due to factors outside of their control (e.g., generational wealth inequality or disability). Allowing factors which are beyond people's control to determine what rights they have may seem unfair.
Desert-Based	Desert is defined in terms of individual effort or contribution, discounting the effects of luck. The amount an individual deserves is thus proportionate to how much they have contributed, after luck has been discounted for. The effects of luck are discounted for because the prior prevalence of a trait in a population can be the result of unjust circumstances (Leben [85]). Dwork et al. [44] suggests desert-based proportionalism could be implemented by assigning each individual some distance in a metric space that evaluates desert, and evaluating fairness through average distance between individuals in the metric space. Persson and Hedlund [106] propose utilising desert to consider how responsibility for ethical AI development should be distributed, assigning responsibility according to the contribution of each individual.	A weakness of this principle is that luck is an abstract concept which is difficult to define, and may vary between contexts. Thus, evaluating which traits should be mitigated for is challenging.

3.2.3 Kantian. Kant [78] argues that ethical principles are derived from the logical structure of action, beginning with distinguishing free action (action for which the agent has reasons) from mere behaviour (Kim et al. [81]). Kant's *categorical imperative* grounds all moral duties as it applies unconditionally to rational agents (categorical), and is a command that could be followed, but might not be (imperative) (Johnson and Cureton [77], Wallach et al. [134]). The categorical imperative entails that a rational agent must believe their reasons for acting are consistent with the assumption that all rational agents to whom the reasons apply could engage in the same actions (also known as the *universal law of nature*) (Boddington [18]). For example, 'do not kill' is a categorical imperative: it is categorical in that if all rational agents committed murder, there would be no rational agents left; it is an imperative as rational agents could kill but should not. Derived from the categorical imperative is the *means-end principle* (also known as the *humanity formula*). The means-end principle denotes that treating other people as a means to an end is immoral (Abney [1], Kumar and Choudhury [83]). It would never be possible to universalise the treatment of another as a means to some end; doing so would contradict the categorical imperative. This contradiction occurs because of our ability to engage in rational self-directed behaviour.

Kantian ethics have been operationalised in previous literature through the imposition of rules. Limarga et al. [89] implement the categorical imperative with two rules: firstly, since it is universal, an agent, in adopting a principle to follow (or judging an action to be its duty), must simulate a world in which everybody abides by that principle and consider that world ideal. Secondly, since actions are inherently morally permissible, forbidden, or obligatory, an agent must perform their duty purely because it is one's duty, and not as a means of achieving an end or by employing another human as a means to an end. Berreby et al. [15] implement the means-end principle in the rule that an action is impermissible if it involves and impacts at least one person, but that impact is not the aim of the action. Svegliato et al. [129] use the moral rule that policies should be universalisable to stakeholders without contradiction. Allen et al. [4] suggest that the categorical imperative could be implemented as a higher principle to evaluate other rules. For example, when deciding whether to apply egalitarianism (ensuring equal distribution), an agent could evaluate if this is the right thing to do by examining if it aligns with the categorical imperative, i.e., if it would be rational for all agents to apply that principle.

A difficulty with the categorical imperative is that it may be too permissive; it could permit intuitively bad things by allowing any action that can have a universalisable maxim (Abney [1]). A common example of this is letting a murderer into your house because you cannot lie, and say that the person they want to kill is not there. The means-end principle can also be too stringent as, interpreted strictly, it forbids any action in which a person affects another without their explicit consent. There are issues that arise related to motivation and free will. According to Kant, the motivation and reasons for why actions are taken are key to whether the action is ethical. However, truly understanding motivation for action may require a level of self-awareness that in practice is difficult to achieve (Boddington [18]). Chakraborty and Bhuyan [26] argue that AI cannot truly implement Kantian ethics without having free will, which is necessary to possess autonomy and the power of reasoning in the Kantian sense.

3.3 Virtue Ethics

According to virtue ethics, ethicality stems from the inherent character of an individual, and not the rightness or wrongness of individual acts (Yu et al. [141]). Right action is performed by someone with virtuous character. In following this theory, one should not be asking what one ought to do, but rather what sort of person one should be (Saltz et al. [113]). The qualities one possesses should be of primary importance, and actions secondary. Moral virtues can be learnt and developed through habit and practice. The stability of virtues (if one has a virtue, one can't behave as if one

doesn't have it) entails that virtue ethics may be a useful way of imbuing machines with ethics (Wallach et al. [134]).

Virtue ethics can be used to formulate ideals for the use of AI, or to create AI which is virtuous itself (Heilinger [66]). To improve ethical use of AI, Robbins and Wallace [110] advocate for applying virtuous characteristics to resolve problems. Vanhée and Borit [131] propose that this may be aided by using education to help designers of systems develop virtues. Hagendorff and Danks [64] advance this by delineating that teaching virtues involves imparting tacit knowledge, social perception skills, and emotion, leading to the automatic 'feeling' of the right thing to do.

Other works focus on implementing virtues directly into machines; according to Tolmeijer et al. [130], inputs for implementing virtue ethics in machines would be properties of the agent, the decision criteria would be based on virtues, and this would be mechanised through the instantiation of virtues. Instantiating virtue ethics in automated decision making is exemplified by Govindarajulu and Bringsjord [59], who define virtues as learnt by experiencing the emotion of admiration when observing virtuous people, and then copying the traits of those people. Computational formal logic is used to formalise emotions (in particular the emotion of admiration), represent (virtuous) traits, and establish a process of learning traits. Greene et al. [61] argue that a virtue-based system would have to appreciate the entire variety of features which call for one action rather than another in a given situation.

Virtue ethics can also be used alongside other approaches; Hagendorff [62] argue deontological approaches should be combined with virtue ethics, using virtues to examine values and character dispositions. Pflanzer et al. [107] suggest implementing virtue ethics to assess the character of an agent in a model which also utilises deontology and consequentialism. Hendrycks et al. [68] implement virtue ethics as part of an assessment criteria, integrating scenarios demonstrating virtue ethics into a dataset used to measure the ethical ability of natural language processing models. Jiang et al. [75] use this dataset to test their model trained on people's moral judgements of various situations.

A problem with virtue ethics highlighted by Saltz et al. [113] is that the holistic view it takes makes it more difficult to apply to individual situations. Tolmeijer et al. [130] identify further challenges relating to the concretion of virtues and conflicting virtues. To judge whether a machine or human is virtuous is not possible by just observing one action or a series of actions that seem to imply virtue—reasons behind actions need to be clear. Requiring understandings of reasons behind actions makes it difficult to build virtues into machines, as there is a high level of abstraction in defining virtues. Additionally, conceptions of virtues can change greatly across time and culture. Virtues instantiated in machines today may lead to unfair outcomes in the future as virtues change, or certain virtues may conflict with other virtues.

3.4 Consequentialism

In consequentialist approaches, right actions are identified through their effects (Brink [23]). The moral validity of an action can thus be judged only by considering its consequences (Rodriguez-Soto et al. [112]). Consequentialist principles can be used to weigh risks and opportunities (Heilinger [66]). A strength of this is that it can be used to evaluate decisions with complex outcomes where some benefit and some are harmed, by examining how benefits and harms are distributed. It can thus explain many moral intuitions that trouble deontological theories, as consequentialists can say that the best outcome is the one in which benefits outweigh costs (Sinnott-Armstrong [121]). In addition, Boddington [18] asserts that the goal-based nature of consequentialist theories suits computing well.

Consequentialist principles can be operationalised by analysing the consequences of different actions. Assessing ethics through consequences denotes a different approach to deontology, which

regards mental states as very important for determining the ethicality of an action. For consequentialism, Tolmeijer et al. [130] denotes that mental states can be largely disregarded. Pflanzer et al. [107] propose a multi-theory model in which consequentialism analyses the consequences brought about by a situation. Deontology and virtue ethics are used in conjunction with consequentialism to make ethical judgements.

Consequentialism has been implemented by weighing actions. Limarga et al. [89] assign each action a weight according to its worst consequence. Actions are part of a sequence to reach a goal, and their weights accumulate to a total amount. This total amount is optimised to select the sequence with the best overall consequence. Suikkanen [126] similarly suggests ranking agents' options in terms of how much aggregate value their consequences have. An option is right if and only if there are no other options with higher evaluative ranking. Tolmeijer et al. [130] argue that input for consequentialist principles would be the action (and its consequences), and the decision criteria would be the comparative well-being. This would be mechanised by selecting the consequence with maximum utility. For binary classification algorithms, Leben [85] suggests implementing consequentialism by examining how weights are assigned to each group outcome based on relative social cost.

However, in practice assigning weights to each outcome may be unrealistic to do for all outcomes (Leben [85]). There might be high computational costs if machines attempt to represent all possible outcomes available (Greene et al. [61]). A related issue is that estimating long-term or uncertain consequences and determining which consequences should be taken into account is difficult (Boddington [18], Etzioni and Etzioni [48]). There may be moral constraints outside consequentialism which prohibit certain actions even when they have the best outcomes, therefore rendering consequentialist theories incomplete (Suikkanen [126]). Another common criticism of consequentialism concerns deciding what is valuable or intrinsically good: whether it is pleasure, preference-satisfaction, the perfection of one's essential capacities, or some list of disparate objective goods (e.g., knowledge, beauty, etc.) (Boddington [18], Brink [23], Tolmeijer et al. [130]).

3.4.1 Utilitarianism. Utilitarianism denotes that the ultimate end is an existence exempt from pain and as rich in enjoyment as possible (Mill [93]). Acts are evaluated by their consequences; an act is ethical if and only if it maximises the total net expected utility across all who are affected (Kim et al. [81]). Requirements for implementing utilitarianism include an account of what outcomes are being aimed for, how to aim for those outcomes, how to measure those outcomes, and what or who matters in assessing and aiming for those outcomes (Boddington [18]).

Utilitarianism has been applied to assess fairness metrics and language models. To justify design choices for fairness metrics in binary classification algorithms, Leben [85] suggests that a function could model each distribution and its effects (a utility function/measure of happiness outcomes); then run a selection procedure over aggregate utilities to maximise the sum. Hendrycks et al. [68] present a dataset of scenarios demonstrating utilitarian principles to analyse the ethical knowledge of natural language processing models. Jiang et al. [75] use this dataset to test their model trained on people's moral judgements of various situations.

Utilitarianism has also been used to select norms which promote value alignment. In MAS, Serramia et al. [116] implement a recursive utility function which identifies the preference utility of each value; the value support of a norm is calculated by adding the utility of each value for that norm. Serramia et al. [117] expands this to assess normative systems. To aggregate value preferences, Lera-Leri et al. [88] implement utilitarianism as a distance function, selecting the optimum from the point of view of the majority. Similarly, Bakker et al. [13] aggregate value preferences estimated by a reward model, implementing utilitarianism to select the maximum mean consensus in the

group. In both Lera-Leri et al. [88] and Bakker et al. [13], social welfare functions are parametric to allow for implementation of different principles (ranging from utilitarian to Rawlsian).

Approaches to operationalise utilitarianism in decision making includes training agents to make judgements that deliver the greatest happiness to the greatest number of people, as in Kumar and Choudhury [83]. Limarga et al. [89] assign a value to every action which is later used for final evaluation. Azad-Manjiri [11] and Dehghani et al. [35] select the choice with the highest utility. In Svegliato et al. [129], autonomous systems make ethically compliant decisions in moral contexts by decoupling the moral principle from the decision module, having a separate moral rule (such as utilitarianism) which evaluates the suggested policy.

A common criticism of utilitarianism is that it could lead to a minority being treated unfairly for the greater good (Anderson et al. [7]). In addition, the theory cannot account for the notion of rights and duties or moral distinctions between, for example, killing versus letting die (Abney [1]). There are also difficulties that arise with quantifying utility. Firstly, calculating the utility of every outcome may be computationally infeasible in scenarios with a very large or infinite number of possible outcomes. Secondly, quantifying utility is difficult as there are different ways of conceptualising what utility means. For instance, whether there is a distinction between higher and lower pleasures will affect how outcomes are quantified (Etzioni and Etzioni [48]). Different qualitative understandings of utility necessitates different ways of quantifying it. To mitigate these issues, utilitarianism could be an additional necessary condition, rather than the sole ethical principle (Kim et al. [81]). Applying utilitarianism as an additional condition would allow for a different ethical principle to provide moral distinctions which are ambiguous in utilitarianism. For sub-types of utilitarianism, see Table 5.

Table 5. Sub-types of Utilitarianism.

Principle	Description	Difficulties
(Hedonic) Act Utilitarianism	Morality of action lies in its consequences (Tolmeijer et al. [130]). Hedonic act utilitarianism entails computing the action which derives the greatest net pleasure (Brink [23]). Berreby et al. [15] suggests a machine utilising this could weigh actions corresponding to their consequences, and then order them accordingly; an action is less desirable if there is another action whose weight is greater. Anderson et al. [7] propose that one could input the number of people affected and the intensity of pleasure/displeasure for each person for each possible action. The algorithm then computes the product of intensity, duration, and probability to obtain the net pleasure for each person. This computation is performed for each alternative action. Nashed et al. [100] implement act utilitarianism by requiring policies which maximise the value of all relevant agents.	A criticism of hedonic act utilitarianism is that it is difficult to define pleasure; what is pleasurable for one person may not be pleasurable for another. Ambiguity in defining pleasure thereby makes it difficult to identify the action with the greatest net pleasure.

Principle	Description	Difficulties
Rule Utilitarianism	Actions are morally assessed by first appraising moral rules based on the principle of utility; deciding whether a (set of) moral rule(s) will lead to the best overall consequences, assuming all/most agents follow it. Berreby et al. [15] illustrate that this could be implemented using a predicate which compounds all effective weights of the actions belonging to a particular rule, then summing up those weights via a predicate. Governatori et al. [58] provide an argumentation framework, where moral theories including rule utilitarianism are expressed as normative systems whose moral justification agents argue about.	Sometimes a rule may lead to unintuitive outcomes, and therefore should be broken. This makes rule utilitarianism look more like act utilitarianism, where the right thing to do is evaluated through the consequences of each action.

3.4.2 Maximin. The maximin principle emphasises maximising the minimum utility by seeking to improve the worst-case experience in a society; guaranteeing a higher than worst-case minimum utility to each individual (Rawls [109]). Maximin thus shifts the focus towards improving the well-being of those who are worst-off (Lee et al. [87]).

Maximin has been implemented in the domain of algorithmic fairness. To evaluate fairness metrics for binary classification, Leben [85] demonstrates how a function modelling each potential distribution and its effects could be constructed, and then a selection procedure run over aggregate utilities. Diana et al. [38] implement maximin to measure fairness by examining worst-case outcomes across all groups, rather than differences between group outcomes. Sun et al. [127] promote fairness by minimising the maximum cost of an allocation over all allocations. Chen and Hooker [27] couple maximin with utilitarianism in optimisation problems to ensure the least advantaged have priority, but not at unlimited cost to everyone else.

Other applications for maximin include preference aggregation. Lera-Leri et al. [88] formulate maximin as a distance function, selecting the optimum solution from the point of view of the most displaced. Bakker et al. [13] estimate preferences in a reward model, and then implement maximin to select the consensus which maximises expected agreement for the most dissenting member. Parametric functions are used to implement different principles such as utilitarianism and maximin in both Bakker et al. [13] and Lera-Leri et al. [88]. Ashrafian [9] proposes implementing maximin using algorithmic game theory to assist governmental policy decisions. Governatori et al. [58] encodes maximin in an argumentation framework for reasoning about different moral theories.

In some situations however, maximin is seen as too risk averse. Consider two situations: A, where there is a 70% chance of gaining £100 and a 30% chance of losing £30; B, where there is a 50% chance of gaining £10 and a 50% chance of losing £10. Sunstein [128] argues maximin would promote choosing option B, but under standard accounts of rationality it would be preferable to choose option A, as the expected value is much higher. Thus, when expected value is high, a reasonable level of risk is preferable to low risk and low expected value. On the other hand, maximin is preferable if we do not know how bad the worse outcome would be (i.e., how much it would decrease welfare, where welfare is not synonymous with expected value), or if it would be catastrophically bad.

3.4.3 Envy-Freeness. In an envy-free allocation, no agent envies another agent (Sun et al. [127]). Fairness thus exists when there are minimal levels of envy between groups or individuals. Resources

may be unequally distributed, but as long as agents do not envy one another, this is considered fair (Boehmer and Niedermeier [19]).

To implement envy-freeness, Boehmer and Niedermeier [19] propose that an assignment of resources to agents is ethical if no agent prefers another agent's bundle (of resources) to their own.

Arguably, what is important might not be a relative condition to other people, but if people have enough to have satisfactory life prospects (Lee et al. [87]). Also, the existence of an envy-free allocation can't be guaranteed when items are indivisible, e.g., chores that need to be assigned to multiple agents. Problems with guaranteeing envy-freeness has led Sun et al. [127] to implement relaxations of the principle, such as envy-free up to one item.

3.4.4 Doctrine of Double Effect. The doctrine of double effect suggests that deliberately inflicting harm is wrong, even if it leads to good (Deng [36]). On the other hand, inflicting harm might be acceptable if it is not deliberate, but simply a consequence of doing good. For this principle, an action is permissible if the action itself is morally good or neutral, some positive consequence is intended, no negative consequence is a means to the goal, and the positive consequences sufficiently outweigh negative ones (Govindarajulu and Bringsjord [59], Lindner et al. [90]).

Govindarajulu and Bringsjord [59] using formal logic to automate the doctrine of double effect, and also the stronger version of the doctrine of triple effect. They use the framework in two different modes: to build doctrine of double effect compliant autonomous systems from scratch, or to verify that a given AI system is doctrine of double effect compliant. Another approach by Berreby et al. [15] implements this principle through rules that proscribe an action if it is intrinsically bad, if it causes a bad effect which leads to a good effect, and if its overall effects are bad.

An issue with the doctrine of double effect is that it still allows bad actions to happen as long as they are not intended, which may have some morally dubious outcomes.

3.4.5 Do No (Instrumental) Harm. People should be free to act as they wish unless doing so would result in harm to another person (Gabriel [56]). Do no instrumental harm allows for harm as a side effect, but not as a means to a goal.

Lindner et al. [90] implements do no harm by stating that a technical agent may not perform an action which causes any harm. Dennis et al. [37] utilise do no harm to ensure agents select plans which can be formally verified as ethical. Alibašić [3] suggests that in the context of cryptocurrency trading, AI should be developed so that it avoids outcomes which cause harm to stakeholders such as individual traders, investors, and the larger community. Harms in this context can occur through different channels such as market manipulation, insider trading, and fraud.

Sometimes, however, there may be situations in which causing harm is inevitable. In such situations, this principle alone would not be able to give clear ethical guidance.

3.5 Other Principles

In addition to the principles mapped out here, there are other principles mentioned in literature which we now describe. For reasons that shall be stated, we did not include these in the taxonomy.

3.5.1 Egoism. Egoism is acting to reach the greatest outcome possible for one's self, irrespective of others (Kumar and Choudhury [83], Robbins and Wallace [110]). Alibašić [3] argues egoism entails assessing if outcomes benefit the interest of the individual or group. In the context of AI and cryptocurrency trading, this would entail selecting outcomes which are better for the system's investors. Elsewhere, this principle is rarely mentioned in literature and this may be because it would lead to likely unethical outcomes if it was imbued in AI agents. If agents were primarily concerned with themselves, irrespective of others, it seems unlikely that they would be ethical

(which involves one party's concern for another, Murukannaiah and Singh [99]). This is because fairness is aimed at the well-being of others as well as the self, whereas egoism is solely self-centred.

3.5.2 Particularism. Particularism emphasises that there is no unique source of normative value, nor is there a single, universally applicable procedure for moral assessment (Tolmeijer et al. [130]). Rules or precedents can guide evaluative practices, however, they are deemed too crude to do justice to many individual situations. Therefore, the moral relevance of a certain feature and the role that it plays will be sensitive to other features of the situation. Ethical evaluation should thus be carried out on a case-by-case basis. Inputs for particularism could include the situation (context, features, intentions, and consequences), with the decision criteria resting on rules of thumb and precedent, as all situations are unique. The mechanism to decide upon an action would depend on how much it fits with rules of thumb or precedents. Jiang et al. [75] present a model to learn descriptive ethics from a data resource of people's ethical judgements of situations. Some challenges identified are that there is no unique and universal logic, thus each situation needs a unique assessment. Particularism is thus hard to generalise and encode in a reproducible way. Bai et al. [12] argue that we cannot avoid choosing some set of principles or rules in developing AI, whether they are implicit or explicit.

3.5.3 The Ethic of Care. This principle emphasises feelings of interconnectedness with others, building on the motivation to look after those who are vulnerable or dependent (Gilligan [57], Robbins and Wallace [110]). Morality is a tool to care for others through nurturing relationships (Kumar and Choudhury [83]). To be ethical, one should think about the situation that others are in. Using your experience, you should act in a nurturing and responsible way. Communication plays an essential role, through the relation of listening and being heard. Care ethics reduces moral distance in AI, where moral distance is when those who are not considered in decisions are treated unethically (Villegas-Galaviz and Martin [133]). Care ethics can be applied to AI by examining the voices not being heard, affected relationships and interdependence, how the system treats context, and if the vulnerable are being exploited (Villegas [132]). Yew [140] advocates for the application of care ethics in the design of robots used for companionship and assistance in healthcare. There is a debate as to whether the ethic of care is a theory in itself, as explored by Held [67], or a practice, virtue, value, or activity which supplements other theories, as Sander-Staudt [114] suggests. Because of this ambiguity we do not include the ethic of care in the taxonomy. However, the ethic of care could be used as a guiding factor in the application of ethical principles, as it enhances the importance of considering others outside of yourself. Emphasis on consideration of others provides good support for value alignment and responsible decision making.

3.5.4 Other Cultures. Lastly, there is a wide variety of principles proposed in cultures outside of the history of Western ethics. Moral frameworks have been established in societies across the world, including Confucian, Shinto, and Hindu thought as well as religious frameworks like Judaism, Christianity, and Islam (Hagerty and Rubinov [65]). There is a multitude of moral frameworks across cultures, with significant variation within these frameworks. Arguably, ethics and culture are inseparable and to understand one you must look at the other. Therefore, ethics must be considered within its cultural context. The reason these principles were not included in the taxonomy is because they would require whole taxonomies of their own. An important direction for future work would be to apply the methodology used in this project specifically to non-Western ethical principles, to form a taxonomy of such principles. Forming taxonomies of principles from a broader variety of cultures will help AI practitioners to build cross-cultural ethical technology.

4 PREVIOUS OPERATIONALISATION OF ETHICAL PRINCIPLES

We iterated over the papers identified in our review to analyse of previous operationalisation of ethical principles for Q_o (Operationalisation). First, we find a variety of technical implementations of ethical principles, summarised in Tables 6 and 7. Second, previous literature integrates principles into reasoning capacities in a top-down, bottom-up, or hybrid architecture, summarised in Table 8. Third, practitioners should be specific about which principle(s) they are operationalising; previous literature suggests that pluralism may help with this decision. Fourth, abstractly, operationalisation falls into the categories of (1) applying rules for deontological principles, (2) developing virtues for virtue ethics, or (3) evaluating consequences for consequentialist principles.

4.1 Choosing Technical Implementation

A variety of technical implementations have been used to encode ethical principles. Expanding upon Tolmeijer et al.'s [130] categorisation, approaches to encode principles into a format computers can understand include logical reasoning, probabilistic reasoning, learning, optimisation, and case-based reasoning. In Table 6 and Table 7, we map each ethical principle found in literature to their technical implementations.

Table 6. Technical implementation of deontological principles and virtue ethics. Papers are categorised by the ethical principles they refer to, and the techniques they employ to implement those principles.

Implementation Type		Ethical Principles				
		Deontology	Egalitarianism	Proportionalism	Kantian	Virtue
Logical Reasoning	Deductive Logic	–	[84]	[84]	[110]	[110]
	Non-Monotonic Logic	–	–	–	[15, 89]	–
	Abductive Logic	–	–	–	–	–
	Deontic Logic	[90]	–	–	–	[60]
	Rule-Based Systems	[30, 35]	[11, 21]	[47]	[110]	[30, 110]
	Event Calculus	–	–	–	[15, 89]	[60]
	Knowledge Representation and Ontologies	[30, 35]	–	–	[110]	[30, 110]
	Inductive Logic	[6, 46]	–	–	[46]	–
Probabilistic Reasoning	Bayesian Approaches	–	–	–	–	–
	Markov Models	[112]	–	–	[112, 129]	[112]

Implementation Type		Ethical Principles				
		Deontology	Egalitarianism	Proportionalism	Kantian	Virtue
	Statistical Inference	–	[44]	[44]	–	–
Learning	Decision Tree	–	[11]	–	–	–
	Reinforcement Learning	[112, 119]	–	–	[112]	[112, 119]
	Inverse Reinforcement Learning	[102]	–	–	–	–
	Neural Networks	[68, 70, 75]	–	–	–	[68, 70, 73, 75]
	Evolutionary Computing	–	–	–	–	[73]
Optimisation		–	[7, 11, 44, 85]	[32, 44, 85]	–	[7]
Case-Based Reasoning		[35, 92]	–	–	–	–

Table 7. Technical implementation of consequentialist principles. Papers are categorised by the ethical principles they refer to, and the techniques they employ to implement those principles.

Implementation Type		Ethical Principles					
		Consequentialism	Utilitarianism	Maximin	Envy-Freeness	Doctrine of Double Effect	Do No Harm
Logical Reasoning	Deductive Logic	–	[84, 110]	–	–	–	
	Non-Monotonic Logic	–	[15, 89]	[15]	–	[15]	–
	Abductive Logic	–	–	–	–	[94]	–
	Deontic Logic	–	[90]	–	–	[59, 90]	[90]
	Rule-Based Systems	[30]	[11, 35, 58, 110]	[2, 58]	–	–	[37]
	Event Calculus	–	[15, 89]	[15]	–	[15, 60]	–
	Knowledge Representation and Ontologies	[30]	[35, 110]	–	–	–	–
	Inductive Logic	–	–	–	–	–	–
Probabilistic Reasoning	Bayesian Approaches	–	[8]	–	–	–	–
	Markov Models	[112]	[100, 129]	–	–	[59]	–
	Statistical Inference	–	–	–	–	–	–

Implementation Type		Ethical Principles					
		Consequentialism	Utilitarianism	Maximin	Envy-Freeness	Doctrine of Double Effect	Do No Harm
Learning	Decision Tree	–	[11]	–	–	–	–
	Reinforcement Learning	[112]	[119]	–	–	–	–
	Inverse Reinforcement Learning	–	–	–	–	–	–
	Neural Networks	–	[13, 68, 70, 75]	[13]	–	–	–
	Evolutionary Computing	–	–	–	–	–	–
Optimisation		–	[7, 8, 11, 27, 85, 88, 117]	[27, 38, 85, 88, 105]	[127]	–	–
Case-Based Reasoning		–	[35]	–	–	[17]	–

4.2 Clarifying the Architecture

To engineer morally sensitive systems, Wallach et al. [134] argue that practitioners must decide on the architecture for integrating ethical principles. These fall within three broad approaches: (1) top-down imposition of ethical theories; (2) bottom-up building of systems with goals that may or may not be explicitly specified; (3) hybrid approaches which combine top-down and bottom-up features. We discuss examples of each architecture and issues which may arise. Table 8 summarises our findings of the technical implementation of ethical principles according to the various architectures.

4.2.1 Bottom-Up Approaches. Bottom-up approaches involve machines learning to make ethical decisions by observing human behaviour in actual situations, without being taught any formal rules or moral philosophy (Etzioni and Etzioni [48]). Bottom-up techniques include artificial neural networks, reinforcement learning, and evolutionary computing (Tolmeijer et al. [130]). An example of this is Noothigattu et al. [102], who use inverse reinforcement learning to align agents with human values by learning policies from observed behaviour. In future work, inverse reinforcement learning could be used to align policies with ethical principles, in a similar way to how Noothigattu et al. [102] align policies with human values. Kim et al. [81] suggest this may improve explainability by assimilating policies with principles which, by their nature, imply logical propositions that can be reasoned about. Dyoub et al. [46] utilise answer set programming (ASP) as a knowledge representation and reasoning language to deductively encode ethical rules. They then utilise inductive logic programming to identify the missing ASP rules needed for ethical reasoning, by learning the relation between the ethical evaluation of an action and related facts in that action's case scenario.

A challenge of bottom-up approaches, however, lies in the risk that machines learn the wrong rules, or cannot reliably extrapolate to cases not reflected in the training data.

4.2.2 Top-Down Approaches. Top-down approaches install ethics directly into the machine, instead of asking the machine to learn from experience, as in bottom-up approaches (Kim et al. [81]).

We find that many works use top-down approaches to integrate ethical principles into reasoning capacities of machines. Dehghani et al. [35] implement deontological and utilitarian principles through a combination of qualitative modelling, first-principles logical reasoning, and analogical reasoning. Tolmeijer et al. [130] found that principles can be implemented as rules through logical or case-based reasoning, using domain knowledge to reason about the situation given as input. Bai et al. [12] do not explicitly encode principles from normative ethics, but provide a methodology in which a set of principles implemented in a top-down fashion forms a 'constitution' and is used to fine-tune a preference model. Parts of their constitution can be aligned to theories like virtue ethics, for example, 'choose the response that a wise, ethical, polite and friendly person would more likely say', where 'wise, ethical, polite and friendly' could be conceptualised as virtues. The preference model is then used to train a reinforcement learning agent.

Top-down approaches have been utilised for optimisation tasks. Serramia et al. [117] implement utilitarianism to optimise for norm systems that promote the most preferred values in a society. Diana et al. [38] operationalise the principle of minimax (minimising the maximum loss, adapted from maximin - maximise the minimum) using oracle-efficient learning algorithms. Minimax is applied to analyse fairness considerations in differences between group outcomes. Also considering fairness, Sun et al. [127] formalise envy-freeness as rules to examine the trade-off between different fairness allocations. Chen and Hooker [27] combine the principles of maximin and utilitarianism in a model for mixed integer and linear programming which can be applied in a top-down manner to optimise social welfare functions. Lera-Leri et al. [88] operationalise different ethical principles by tuning the parameter of a function, which is then applied as a distance function to optimise value preference aggregation.

However, as human knowledge does not tend to be very structured, domain knowledge needs to be interpreted before it can be used. A difficulty of top-down approaches is that human understandings of philosophical rules need to be encoded in a way that machines can understand, which may mean that information is lost or misrepresented.

4.2.3 Hybrid Approaches. Hybrid approaches embody aspects of both top-down and bottom-up approaches. As top-down and bottom-up approaches each employ different aspects of moral sensibility, combining the two may result in better implementation of ethical principles (Allen et al. [4]). A benefit of hybrid approaches is that they incorporate both ethical reasoning and empirical observation, which allows context to be taken into account.

Hybrid architectures have been used in individual decision making through logic and reinforcement learning. Berreby et al. [15] supplement top-down imposition of rules with bottom-up observation of contextual information, allowing agents to represent and reason about a variety of deontological and consequentialist theories. They propose a modular logic-based framework based on a modified version of the Event Calculus, implemented in Answer Set Programming. Limarga et al. [89] implement principles using non-monotonic reasoning in an event set calculus, which allows rules to be revised when a conflict arises. Rodriguez-Soto et al. [112] provide a method that first characterises ethical behaviour as ethical rewards, and then embeds such rewards into the learning environment of the agent using multi-objective reinforcement learning. Following a top-down approach, ethical principles are formalised along normative (whether the action is good or bad) and evaluative dimensions (how good it is). In a bottom-up manner, the principles are then used as reward functions.

For optimisation, Bakker et al. [13] operationalise ethical principles by tuning the parameter of a function, which is then used to aggregate preferences estimated by a reward model.

Hybrid architectures have also been used in the context of large language models (LLMs). Hendrycks et al. [68] construct a dataset of contextualised scenarios demonstrating a variety of

ethical principles to assess ethical knowledge learnt by LLMs. Jiang et al. [75] present a data resource of people's judgements of ethical situations, and use it to train a model. The authors test the model against tasks implementing ethical principles from Hendrycks et al.'s [68] dataset. Shi et al. [119] implement Hendrycks et al. [68] in a plugin moral-aware learning model to train a reinforcement learning agent which alternates between learning tasks and morality in a text-based environment.

Whilst there are benefits from combining aspects of top-down and bottom-up architectures, there difficulties also emerge from the meshing of dissimilar architectures and diverse ideas about the origins of morality. Where top-down approaches emphasise ethical concerns arising from outside the entity, bottom-up approaches focus on ethics arising within, embodying different aspects of moral sensibility. Hybrid systems must be able to balance tensions between internal and external ethical concerns (Allen et al. [4], Wallach et al. [134]).

Table 8. Architecture for implementing principles. Papers are categorised by principles they refer to and the architecture used, where bottom-up involves machines learning to make ethical decisions through observation; top-down involves imposition of rules; hybrid involve a combination of bottom-up and top-down techniques.

Ethical Principles	Bottom-Up	Top-Down	Hybrid
Deontology	Inductive Logic [6] Inverse Reinforcement Learning [102]	Rule Based Systems, Knowledge Representation and Ontologies, Case-Based Reasoning [35] Deontic Logic [90] Case Based Reasoning [92]	Neural Networks [68, 70, 75] Rule-Base Systems, Knowledge Representation and Ontologies [30] Markov Models, Reinforcement Learning [112] Reinforcement Learning [119]
Egalitarianism	–	Deductive Logic [84] Statistical Inference, Optimisation [44] Optimisation [7, 85] Rule-Based Systems [21]	Rule-Based Systems, Decision Tree, Optimisation [11]
Proportionalism	–	Deductive Logic [84] Statistical Inference, Optimisation [44] Rule-Based Systems [47] Optimisation [32, 85]	–
Kantian	–	Deductive Logic, Rule-Based Systems, Knowledge Representation and Ontologies [110] Markov Models [129]	Non-Monotonic Reasoning and Event Calculus [15, 89] Markov Models, Reinforcement Learning [112]

Ethical Principles	Bottom-Up	Top-Down	Hybrid
Virtue	Evolutionary Computing [73]	Deductive Logic, Rule-Based Systems, Knowledge Representation and Ontologies [110]	Neural Networks [68, 70, 75] Deontic Logic, Event Calculus [60] Rule-Base Systems, Knowledge Representation and Ontologies [30] Markov Models, Reinforcement Learning [112] Reinforcement Learning [119]
Consequentialism	–	–	Rule-Base Systems, Knowledge Representation and Ontologies [30] Markov Models, Reinforcement Learning [112]
Utilitarianism	Rule Based Systems, Knowledge Representation and Ontologies [35]	Deductive Logic [84] Deductive Logic, Rule-Based Systems, Knowledge Representation and Ontologies [110] Deontic Logic [90] Optimisation [7, 27, 85, 88, 117] Markov Models [100] Rule-Based Systems [58]	Non-Monotonic Reasoning and Event Calculus [15, 89] Neural Networks [13, 68, 70, 75] Rule-Based Systems, Decision Tree, Optimisation [11] Bayesian Approaches, Optimisation [8] Reinforcement Learning [119]
Maximin	–	Optimisation [27, 38, 85, 88, 105] Rule-Based Systems [2, 58]	Non-Monotonic Reasoning and Event Calculus [15] Neural Networks [13]
Envy-Freeness	–	Optimisation [127]	
Doctrine of Double Effect	–	Abductive Logic [94] Deontic Logic [90] Deontic Logic, Event Calculus, Markov Models [59] Case Based Reasoning [17]	Non-Monotonic Reasoning and Event Calculus [15]
Do No Harm	–	Deontic Logic [90] Rule-Based Systems [37]	–

4.3 Specifying the Ethical Principle

Practitioners should specify which ethical principle(s) will be operationalised. This could be aided by referring to the taxonomy we have suggested, which contains a broad array of ethical principles found in AI and computer science literature (Figure 1). Leben [85] emphasises that being clear about which principle is being used will help designers to further clarify what inputs are necessary for their application, which in turn will improve ethical reasoning capabilities and explainability of how decisions have been made.

4.3.1 Implementing Pluralism. Human morality is complex and cannot be captured by a single classical ethical theory (Robbins and Wallace [110]). Thus, it may not always be easy to decide which principle to apply. Pluralism advocates that there is not one approach that is best. In a similar way to how we learn and implement different programming languages, Brennan [22] argues that we utilise different ethical principles depending on the problem at hand. Context and various reasoning techniques could be used to choose between appropriate principles. Tolmeijer et al. [130] advocate for further research according to this approach, suggesting the development of multi-theory models where machines interchangeably apply different theories depending on the situation.

Pluralism has been operationalised in previous literature. Svegliato et al. [129] propose a framework which decouples ethical compliance from task completion to avoid unanticipated scenarios which do not reflect stakeholder values. They suggest implementing a pluralist approach in the form of an extra moral constraint representing a moral principle. This allows for the decision-making module's policy to be evaluated considering its ethical context, leaving room to implement different ethical principles as the ethical rule. Lera-Leri et al. [88] implement a range of ethical principles as distance functions, and use these functions to aggregate value preferences. Pflanzer et al. [107] propose utilising the Agent-Deed-Consequence model for ethical decision making in AI, which implements virtue ethics to evaluate the character of a person (Agent), deontology to examine their actions (Deed), and consequentialism to assess the consequences brought about by the situation (Consequence). If all components are positive, the moral judgement is positive.

4.4 Choosing Abstract Implementation

We have found that abstract implementation of principles falls into three main categories: rules, consequences, or virtues. We discuss examples of each implementation category and potential difficulties that may arise. Deontological principles have been operationalised by applying rules, and choosing an action based on how it accords with those rules. Virtue ethics has been operationalised by developing virtuous characteristics. Consequentialist principles have been operationalised by evaluating consequences and choosing an action based on the consequences it produces.

4.4.1 Applying Rules. For deontological principles, some approaches suggest operationalising principles by applying a set of rules to possible actions to determine which ones would be satisfactory, such as Abney [1], Berreby et al. [15], and Greene et al. [61]. Examples of this, as suggested by Murukannaiah et al. [98], would be applying the rule that the disparity of preference satisfaction for stakeholders should be minimised, extracted from the principle of egalitarianism. Another example is Leben [85], applying the rule that stakeholders should be treated proportionally based on their contributions to production.

Due to the abstract nature of ethics, difficulties arise in finding appropriate ways to encode ethical principles in concrete rules. One difficulty lies in deciding if rules should be interpreted as strict or defeasible (Tolmeijer et al. [130]). For example, an essential part of Kant's [78] ethics is that the reasons for actions must be universalisable to all agents. The need for reasons to be universal implies that this rule should be strict. However, this could permit actions that are bad according to

other principles, suggesting that it should be defeasible (Abney [1]). Nashed et al. [101] argue that although implementing ethics through rules sets a high standard for agent behaviour, expressive, effective, and general rule sets are difficult to generate. Creating systematic ways of encoding the ethical principles we identify (Figure 1) into rules, including understanding whether rules should be strict or defeasible, to use in the reasoning capacities of AI could thus be a direction for future research.

4.4.2 Developing Virtues. For virtue ethics, ethicality stems from the inherent character of an individual (Kazim and Koshiyama [79]). To solve a problem according to this theory, virtuous characteristics should be applied (Robbins and Wallace [110]). Thus, the theory can be operationalised by instantiating virtues (Tolmeijer et al. [130]). Instantiating virtues is exemplified by Govindarajulu and Bringsjord [59], who understand virtues as learnt by experiencing the emotion of admiration when observing virtuous people, and then copying the traits of those people. This is implemented using computational formal logic to formalise emotions (in particular, the emotion of admiration), represent traits, and establish a process of learning traits. To formalise virtues, the authors use a deontic cognitive event calculus, which is a quantified multi-operator modal logic that includes the event calculus for reasoning over time and change. By formalising emotions (admiration) in this way, agents associate admiration with the actions of others. Traits are formalised as a series of instantiations of a type of behaviour. If enough admiration is felt for particular traits, the agents learn the traits, thus instantiating virtues.

However, virtue ethics can be difficult to apply to individual situations (Saltz et al. [113]), and there are challenges that arise with the application of virtues across time and culture (Tolmeijer et al. [130]). Future research could therefore examine the applicability and appropriateness of virtue ethics across different contexts.

4.4.3 Evaluating Consequences. Consequentialist principles may be operationalised by evaluating the consequences of different actions (Limarga et al. [89]). Suikkanen [126] suggests this could be done by ranking agents' options in terms of how much aggregate welfare their consequences have. Dehghani et al. [35] specify this with the principle of utilitarianism, by selecting the choice with the highest utility. Ajmeri et al. [2] operationalise the principle of maximin by improving the minimum experience in the consequences of an action. Consequences are also used to operationalise the principle of envy-freeness, which Sun et al. [127] address by promoting the outcome with the lowest levels of envy between groups or individuals.

Issues arise in predicting all of the possibilities an action could produce. Predicting all possibilities could be computationally challenging, requiring complex calculations (Greene et al. [61]). There are thus limitations to simulating all possible consequences of an action in non-deterministic and probabilistic environments; future work could explore applying multiple ethical principles to such environments.

5 GAPS IN OPERATIONALISING ETHICAL PRINCIPLES

To address Q_g (Gaps), we now examine existing gaps in ethics and fairness research in AI and computer science literature, specifically in relation to implementing multiple ethical principles in reasoning capacities.

5.1 Expanding the Taxonomy

Understanding strengths and weaknesses of various approaches improves critical understanding and constructive engagement (Boddington [18]). Therefore, key gaps include research on lesser-utilised principles. We suggest that future directions consider less commonly seen principles, or incorporate a wider array of principles. This includes researching principles from other cultures

outside of the Western doctrine, which is important as ethics is culturally sensitive (Hickok [69]). Implementing ethical principles from various cultures will aid the accessibility and fairness of AI, as it can better apply to stakeholders from diverse backgrounds. Hongladarom and Bandasak [71] survey non-western guidelines for AI principles, finding unique cultural presuppositions in some areas and global consensus in others. Expanding this research to examine AI and non-western principles from normative ethics is a future research direction.

5.2 Resolving Ethical Dilemmas

We identify various difficulties with the implementation of ethical principles that may result in ethical dilemmas, from which various gaps arise. Anderson and Anderson [5] define ethical dilemmas as situations where either there is not a good choice between different outcomes, or where the choice between different outcomes is not obvious (e.g., the distinction of how good one outcome is compared to another is not obvious). Future research could address gaps that arise in resolving these dilemmas.

In certain situations, dilemmas arise when the application of one principle cannot support one action over another. Azad-Manjiri [11] suggest that one way to resolve this could be by examining how similar decisions were made previously. If no similar decisions have been made previously, an action is selected at random. However, this approach may run into the naturalistic fallacy, looking at what is the case rather than what ought to be the case. In addition, relying on random choice may not result in the most ethically appropriate action. A gap exists in further examining how to resolve dilemmas where principles cannot support one action of another.

For each principle, dilemmas arise when its application leads to an unfair outcome, as all moral theories have some counter-intuitive implications (Robinson [111]). This implies no single theory can denote how to program ethical AI (Pagallo [104]). Pluralist approaches, in which different principles can be weighed against one another to find the most appropriate answer, could help mitigate these issues. Works such as Governatori et al. [58], Lera-Leri et al. [88], and Pflanzer et al. [107] provide methodologies accommodating multiple principles. A gap exists in applying such methodologies to compare the application of multiple principles in scenarios where particular principles lead to unfair outcomes. However, weighing alternatives may not always be possible. Future research should investigate the feasibility of applying different principles in diverse scenarios.

Dilemmas may arise with the application of multiple principles, as different principles can give different answers which may conflict (Persson and Hedlund [106]). This is exemplified in Nashed et al. [100], who find that agents implementing different principles favour different policies. In addition, it is difficult to apply abstract theories to concrete situations. To aid this, Tolmeijer et al. [130] suggest that particularism (which incorporates relevant contextual factors in ethical reasoning to identify if a certain feature is morally relevant or not) could help identify which principle is the most appropriate in that setting. A gap exists in exploring if aspects of particularism can be used to resolve dilemmas where different principles promote conflicting outcomes. However, there are also issues that arise with the application of particularism. While ethics examines principles that socially impose what's right or wrong, morality deals with social values of right or wrong (Jiang et al. [75]). Moral disagreement can arise when stakeholders have different beliefs about which facts are morally relevant, and which ethical principle is true (Awad et al. [10], Robinson [111]).

There are thus various gaps and difficulties which arise in regard to resolving ethical dilemmas. Drawing these ideas together, there are gaps in finding reliable methodologies for AI practitioners to decide which principle is most appropriate for a particular case, considering the dilemmas which may arise. Robinson [111] explores three solutions which may help to resolve dilemmas: moral solutions, compromise solutions, and epistemic solutions. Moral solutions select a moral theory either by what we think is true, some general theory which we can agree on, or what we could

hypothetically agree on under certain disagreements. Compromise solutions choose principles based on a social choice approach, or treat principle selection as a multi-objective optimisation problem, optimising based on inferred moral values or goals. Epistemic solutions harness information about the disagreement as evidence of moral facts, and then appeal to a rule for decision making under moral uncertainty. Alternatively, epistemic solutions could attempt to achieve an overarching moral view which accommodates as many relevant ethical judgements as possible. Each approach has various strengths and limitations, as discussed in the paper. Robinson [111] concludes that problems of moral disagreement should be treated as problems of managing moral risk, where moral risk is the chance of getting things wrong and what you thereby risk. A gap exists in implementing such solutions to evaluate how they address ethical dilemmas in practice.

5.3 Implementing Ethical Principles in STS

An application of implementing ethical principles in STS is to support governance capacities. Governance of STS involves establishing standards for the proper use and development of technology, as defined by Floridi [52], and administration of systems by stakeholders themselves, as defined by Singh [120]. Under the perspective of macro ethics, responsible governance should incorporate norms and value preferences of different stakeholders. However, dilemmas arise when norms or values conflict. Operationalising ethical principles in reasoning helps support governance capacities to resolve these dilemmas in equitable ways that support the needs of different stakeholders (Woodgate and Ajmeri [138]). Gaps exist relating to how ethical principles can be operationalised in STS to promote equitable governance capacities.

Previous work provides guidance for applying ethical principles to reason about values and norms in computational decision making. For example, Ajmeri et al. [2] broadly reference the principles of egalitarianism and utilitarianism within the context of utilising values and norms in MAS for ethical reasoning in individual decision making. This research may benefit from the consideration of other ethical principles to enable broader applicability. Lera-Leri et al. [88] present a method for applying multiple ethical principles to aggregate different value preferences, but do not consider the influence of norms. Serramia et al. [117] demonstrate how to select norms that best align with a known value system. Combining these approaches to aggregate different value systems using ethical principles, and then using the aggregated value systems to select value-aligned norms, is a promising direction for future research. In addition, a gap exists in examining how such approaches can be incorporated in decentralised collective decision making.

However, challenges arise when implementing ethical principles in STS considering value systems. Norms and values are interdependent with context and decision-makers, and research should consider how value systems change according to context, for example, over time (Osman and d'Inverno [103], Smit and Pitt [122]). Gaps exist related to implementing principles in STS in ways that account for the relationship between context and changing value systems.

Properly incorporating broad social context requires careful consideration to avoid entrenching dominant relations of power (Weinberg [136]). This includes accounting for the ways in which existing dynamics shape how technology is developed and deployed. Applying ethical principles must therefore integrate broad social dynamics, and appreciate how social dynamics affect governance capacities (Munn [97]). Gaps emerge with respect to accommodating broad social dynamics and avoiding perpetuating unjust power dynamics in the application of ethical principles to governance capacities.

To understand how ethical principles can accommodate for broad social dynamics, participatory approaches may be useful to incorporate human input throughout the design process. For example, Dubljević et al. [43] combine participatory approaches with multi-criteria decision making to capture the importance of different harms and make clear the perspectives of different stakeholders.

Weinberg [136] emphasises that collaborating with those affected by the technology improves the propensity to leverage knowledge from marginalised groups, understand how the technology is situated in its social context, and address what is most ethically concerning, rather than what is most convenient to measure. Participatory approaches present opportunities to investigate questions related to what extent ethical principles are generalisable across different groups of people, what people morally disagree on, what preferences people have over ethical principles, and if and how people follow ethical principles in their daily lives. Gaps exist in further examining these questions.

6 CONCLUSION

To better address the pursuit of responsible AI, research must be human-centred (Collins et al. [31], Dignum and Dignum [41]). Shifting the perspective to the macro ethics of STS, considering the range of relevant human values and ethical features, may help to enable responsible ethical-decision making which can be justified and held accountable (Chopra and Singh [29], Lechterman [86]). However, dilemmas arise when values conflict (Murukannaiah et al. [98]). To resolve these dilemmas in satisfactory ways, ethical principles can help to determine the moral permissibility of actions (Lindner et al. [90], McLaren [92]).

We identify a variety of ethical principles which have been previously operationalised in AI and computer science literature. We also identify key aspects of operationalising ethical principles in AI, including selecting technical implementation, clarifying the architecture, specifying the ethical principle, and using rules, consequences or virtues. Key gaps that imply future research directions include expanding the taxonomy, resolving ethical dilemmas where principles conflict or lead to unfair outcomes, and implementing principles in STS whilst accommodating for changing contexts and broad social dynamics. We envision that our findings will contribute towards developing responsible AI by aiding the incorporation of ethical principles in reasoning capacities.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their careful reading and insightful comments which helped us to substantially improve the manuscript. JW thanks the EPSRC Doctoral Training Partnership Grant No. EP/W524414/1 for support. NA acknowledges partial support from the UKRI EPSRC Grant No. EP/Y028392/1: *AI for Collective Intelligence (AI4CI)*.

REFERENCES

[1] Keith Abney. 2011. *Robots, Ethical Theory, and Metaethics: A Guide for the Perplexed.* MIT Press, Cambridge, 35–52.

[2] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Auckland, 16–24. https://doi.org/10.5555/3398761.3398769

[3] Haris Alibašić. 2023. Developing an Ethical Framework for Responsible Artificial Intelligence (AI) and Machine Learning (ML) Applications in Cryptocurrency Trading: A Consequentialism Ethics Analysis. *FinTech* 2, 3 (2023), 430–443. https://doi.org/10.3390/fintech2030024

[4] Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology* 7, 3 (Sept. 2005), 149–155. https://doi.org/10.1007/s10676-006-0004-4

[5] Michael Anderson and Susan L. Anderson. 2007. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28, 4 (Dec. 2007), 15. https://doi.org/10.1609/aimag.v28i4.2065

[6] Michael Anderson and Susan Leigh Anderson. 2014. GenEth: A General Ethical Dilemma Analyzer. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 1. Association for the Advancement of Artificial Intelligence, Québec, 253–261.

[7] Michael Anderson, Susan Leigh Anderson, and Chris Armen. 2004. Towards Machine Ethics. In *AAAI-04 Workshop on Agent Organizations: Theory and Practice*. AAAI, San Jose, 1–7.

[8] Stuart Armstrong. 2015. Motivated Value Selection for Artificial Agents.. In *Proceedings of AAAI Workshop on AI and Ethics*. Austin.

[9] Hutan Ashrafian. 2023. Engineering a social contract: Rawlsian distributive justice through algorithmic game theory and artificial intelligence. *AI and Ethics* 3, 4 (Nov. 2023), 1447–1454. https://doi.org/10.1007/s43681-022-00253-6

[10] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M. J. Crockett, Jim A. C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, Matthew S. Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. 2022. Computational ethics. *Trends in cognitive sciences* 26 (March 2022), 388–405. Issue 5. https://doi.org/10.1016/j.tics.2022.02.009

[11] Meisam Azad-Manjiri. 2014. A New Architecture for Making Moral Agents Based on C4.5 Decision Tree Algorithm. *International Journal of Information Technology and Computer Science* 6, 5 (2014), 50–57. https://doi.org/10.5815/ijitcs.2014.05.07

[12] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL]

[13] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 35. Curran Associates, Inc, New Orleans, 38176–38189. https://proceedings.neurips.cc/paper_files/paper/2022/file/f978c8f3b5f399cae464e85f72e28503-Paper-Conference.pdf

[14] Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. *An Introduction to Ethics in Robotics and AI.* Springer, New York. https://doi.org/10.1007/978-3-030-51110-4

[15] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2017. A Declarative Modular Framework for Representing and Applying Ethical Principles. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, São Paulo, 96–104. https://doi.org/10.5555/3091125.3091145

[16] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAccT)*, Vol. 81. PMLR, New York, 149–159.

[17] Joseph Blass and Kenneth Forbus. 2015. Moral Decision-Making by Analogy: Generalizations versus Exemplars. *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)* 29, 1 (Feb. 2015), 501 – 507. https://doi.org/10.1609/aaai.v29i1.9226

[18] Paula Boddington. 2023. *Normative Ethical Theory and AI Ethics.* Springer Nature Singapore, Singapore, 229–276. https://doi.org/10.1007/978-981-19-9382-4_6

[19] Niclas Boehmer and Rolf Niedermeier. 2021. Broadening the Research Agenda for Computational Social Choice: Multiple Preference Profiles and Multiple Solutions. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Virtual Event, London, 1–5. https://doi.org/10.5555/3463952.3463954

[20] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576. https://doi.org/10.1126/science.aaf2654

[21] Sirin Botan, Ronald de Haan, Marija Slavkovik, and Zoi Terzopoulou. 2023. Egalitarian judgment aggregation. *Autonomous Agents and Multi-Agent Systems* 37, 1 (Feb. 2023), 16. https://doi.org/10.1007/s10458-023-09598-6

[22] Jason Brennan. 2007. *The best moral theory ever: The merits and methodology of moral theorizing.* Ph. D. Dissertation. University of Arizona.

[23] David Brink. 2007. Some Forms and Limits of Consequentialism. *The Oxford Handbook of Ethical Theory* 1, 1 (June 2007), 381–423. https://doi.org/10.1093/oxfordhb/9780195325911.003.0015

[24] Emanuelle Burton, Judy Goldsmith, Sven Koenig, Benjamin Kuipers, Nicholas Mattei, and Toby Walsh. 2017. Ethical Considerations in Artificial Intelligence Courses. *AI Magazine* 38, 2 (July 2017), 22–34. https://doi.org/10.1609/aimag.v38i2.2731

[25] Cansu Canca. 2020. Operationalizing AI Ethics Principles. *Commun. ACM* 63, 12 (Nov. 2020), 18–21. https://doi.org/10.1145/3430368

[26] Arunima Chakraborty and Nisigandha Bhuyan. 2023. Can artificial intelligence be a Kantian moral agent? On moral autonomy of AI system. *AI and Ethics* 3, 2 (March 2023), 1–7. https://doi.org/10.1007/s43681-023-00269-6

[27] Violet (Xinying) Chen and J. N. Hooker. 2020. A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism. In *Proceedings of the 3rd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, New York, 221–227. https://doi.org/10.1145/3375627.3375844

[28] Lu Cheng, K. Varshney, and Huan Liu. 2021. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *JAIR* 71 (Aug. 2021), 1137–1181. https://doi.org/10.1613/jair.1.128142

[29] Amit Chopra and Munindar Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 1st AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, New Orleans, 48–53. https://doi.org/10.1145/3278721.3278740

[30] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgement of Agents' Behaviours in Multi-Agent Systems. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Singapore, 1106–1114. https://hal-emse.ccsd.cnrs.fr/emse-01317409

[31] Daniel E. Collins, Conor J. Houghton, and Nirav Ajmeri. 2024. Fostering Multi-Agent Cooperation through Implicit Responsibility. In *Proceedings of the 2nd International Workshop on Citizen-Centric Multiagent Systems (CMAS)*. Auckland, 1–10.

[32] Vincent Conitzer, Walter Sinnott-Armstrong, J. S. Borg, Yuan Deng, and Max Kramer. 2017. Moral Decision Making Frameworks for Artificial Intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, Honolulu, 4831–4835.

[33] Mehdi Dastani and Vahid Yazdanpanah. 2022. Responsibility of AI Systems. *AI and Society* 1, 1435-5655 (June 2022), 1–10. https://doi.org/10.1007/s00146-022-01481-4

[34] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. 2013. No Smoking Here: Values, Norms and Culture in Multi-Agent Systems. *Artificial Intelligence and Law* 21, 1 (01 March 2013), 79–107. https://doi.org/10.1007/s10506-012-9128-5

[35] Morteza Dehghani, Emmett Tomai, and Matthew Klenk. 2008. An Integrated Reasoning Approach to Moral Decision-Making. *Machine Ethics* 3 (Jan. 2008), 1280–1286. https://doi.org/10.1017/CBO9780511978036.024

[36] Boer Deng. 2015. Machine ethics: The robot's dilemma. *Nature* 523 (July 2015), 24–26. Issue 7558. https://doi.org/10.1038/523024a

[37] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14. https://doi.org/10.1016/j.robot.2015.11.012

[38] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Convergent Algorithms for (Relaxed) Minimax Fairness. *CoRR* abs/2011.03108 (2021), 1–22. arXiv:2011.03108 [cs.LG]

[39] Virginia Dignum. 2017. Responsible Artificial Intelligence: Designing AI for Human Values. *ITU Journal* 1, 1 (2017), 1–8. Issue 1. https://api.semanticscholar.org/CorpusID:158378636

[40] Virginia Dignum. 2019. *Ethical Decision-Making*. Springer, Cham, 35–46. https://doi.org/10.1007/978-3-030-30371-6_3

[41] Virginia Dignum and Frank Dignum. 2020. Agents Are Dead. Long Live Agents!. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Virtual Event, New Zealand, 1701–1705. https://doi.org/10.5555/3398761.3398957

[42] Veljko Dubljević, Sean Douglas, Jovan Milojevich, Nirav Ajmeri, William A. Bauer, George F. List, and Munindar P. Singh. 2021. Moral and Social Ramifications of Autonomous Vehicles. *CoRR* abs/2101.11775 (Jan. 2021), 1–8. arXiv:2101.11775

[43] Veljko Dubljević, George List, Jovan Milojevich, Nirav Ajmeri, William A. Bauer, Munindar P. Singh, Eleni Bardaka, Thomas A. Birkland, Charles H. W. Edwards, Roger C. Mayer, Ioan Muntean, Thomas M. Powers, Hesham A. Rakha, Vance A. Ricks, and M. Shoaib Samandar. 2021. Toward a rational and ethical sociotechnical system of autonomous vehicles: A novel application of multi-criteria decision analysis. *PLOS ONE* 16, 8 (Aug. 2021), 1–17. https://doi.org/10.1371/journal.pone.0256224

[44] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. ACM, Cambridge, 214–226.

[45] Ronald Dworkin. 1981. What is Equality? Part 1: Equality of Welfare. *Philosophy and Public Affairs* 10, 3 (1981), 185–246. https://doi.org/10.2307/2264894

[46] Abeer Dyoub, Stefania Costantini, and Francesca Alessandra Lisi. 2022. Learning Domain Ethical Principles from Interactions with Users. *Digital Society* 1 (Nov. 2022), 28. https://doi.org/10.1007/s44206-022-00026-y

[47] Amitai Etzioni and Oren Etzioni. 2016. AI assisted ethics. *Ethics and Information Technology* 18 (June 2016), 149–156. Issue 2. https://doi.org/10.1007/s10676-016-9400-6

[48] Amitai Etzioni and Oren Etzioni. 2017. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21 (Dec. 2017), 403–418. Issue 4. https://doi.org/10.1007/s10892-017-9252-2

[49] Sina Fazelpour, Zachary C. Lipton, and David Danks. 2022. Algorithmic Fairness and the Situated Dynamics of Justice. *Canadian Journal of Philosophy* 52, 1 (2022), 44–60. https://doi.org/10.1017/can.2021.24

[50] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. In *Berkman Klein Center Research Publication No. 2020-1*. Berkman Klein Center, Cambridge, 1–39. https://doi.org/10.2139/ssrn.3518482

[51] Marc Fleurbaey. 2008. *Fairness, Responsibility, and Welfare*. Oxford University Press, Oxford.

[52] Luciano Floridi. 2018. Soft Ethics and the Governance of the Digital. *Philosophy & Technology* 31 (2018), 1–8. Issue 1. https://doi.org/10.1007/s13347-018-0303-9

[53] Luciano Floridi. 2018. Soft Ethics: Its Application to the General Data Protection Regulation and Its Dual Advantage. *Philosophy & Technology* 31, 2 (June 2018), 163–167. https://doi.org/10.1007/s13347-018-0315-5

[54] Luciano Floridi and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1, 1 (July 2019), 1. https://doi.org/10.1162/99608f92.8cd550d1 https://hdsr.mitpress.mit.edu/pub/l0jsh9d1.

[55] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM (CACM)* 64, 4 (March 2021), 136–143. https://doi.org/10.1145/3433949

[56] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (Sept. 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[57] Carol Gilligan. 1993. *In a Different Voice: Psychological Theory and Women's Development.* Harvard University Press, Harvard. http://www.jstor.org/stable/j.ctvjk2wr9

[58] Guido Governatori, Francesco Olivieri, Regis Riveret, Antonino Rotolo, and Serena Villata. 2018. Dialogues on moral theories. In *Deontic Logic in Computer Science (DEON '18)*. College Publications, Utrecht.

[59] Naveen Sundar Govindarajulu and Selmer Bringsjord. 2017. On Automating the Doctrine of Double Effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, Melbourne, 4722–4730. https://doi.org/10.24963/ijcai.2017/658

[60] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the Engineering of Virtuous Machines. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, Honolulu, 29–35. https://doi.org/10.1145/3306618.3314256

[61] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. 2016. Embedding Ethical Principles in Collective Decision Support Systems. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, Utah, 4147–4151. https://doi.org/10.5555/3016387.3016503

[62] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30 (March 2020), 99–120. Issue 1. https://doi.org/10.1007/s11023-020-09517-8

[63] Thilo Hagendorff. 2022. A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philosophy & Technology* 35, 3 (June 2022), 55. https://doi.org/10.1007/s13347-022-00553-z

[64] Thilo Hagendorff and David Danks. 2023. Ethical and methodological challenges in building morally informed AI systems. *AI and Ethics* 3, 2 (May 2023), 553–566. https://doi.org/10.1007/s43681-022-00188-y

[65] Alexa Hagerty and Igor Rubinov. 2019. Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. *CoRR* abs/1907.07892 (July 2019), 1–27. arXiv:1907.07892

[66] Jan-Christoph Heilinger. 2022. The Ethics of AI Ethics. A Constructive Critique. *Philosophy & Technology* 35, 3 (July 2022), 61. https://doi.org/10.1007/s13347-022-00557-9

[67] Virginia Held. 2005. 9 The Ethics of Care as Moral Theory. In *The Ethics of Care: Personal, Political, and Global*. Oxford University Press, Oxford. https://doi.org/10.1093/0195180992.003.0002

[68] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI With Shared Human Values. *CoRR* abs/2008.02275 (2020), 1–29. arXiv:2008.02275

[69] Merve Hickok. 2021. Lessons learned from AI ethics principles for future actions. *AI and Ethics* 1 (2021), 41–47. Issue 1. https://doi.org/10.1007/s43681-020-00008-1

[70] Ali Reza Honarvar and Nasser Ghasem-Aghaee. 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)*. IEEE, Daejeon, 290–295. https://doi.org/10.1109/CIRA.2009.5423190

[71] Soraj Hongladarom and Jerd Bandasak. 2023. Non-western AI ethics guidelines: implications for intercultural ethics of technology. *AI and Society* 38 (April 2023), 1–14. Issue 2. https://doi.org/10.1007/s00146-023-01665-6

[72] Oscar Horta, Gary David O'Brien, and Dayron Teran. 2022. The Definition of Consequentialism: A Survey. *Utilitas* 34, 4 (2022), 368–385. https://doi.org/10.1017/S0953820822000164

[73] Muntean Ioan and Don Howard. 2017. Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency. In *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*, Thomas Powers (Ed.). Springer, Online.

[74] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, Seoul, 310–323. https://doi.org/10.1145/3531146.3533097

[75] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jonathan Borchardt, Jenny T. Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards Machine Ethics and Norms. *CoRR* abs/2110.07574 (2021), 1–42. arXiv:2110.07574

[76] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2
[77] Robert Johnson and Adam Cureton. 2022. Kant's Moral Philosophy. In *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University, Stanford.
[78] Immanuel Kant. 2011. *Immanuel Kant: Groundwork of the Metaphysics of Morals: A German-English edition.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511973741
[79] Emre Kazim and Adriano Koshiyama. 2020. A High-Level Overview of AI Ethics. *SSRN* 1, 1 (May 2020), 1–18. https://doi.org/10.2139/ssrn.3609292
[80] Arif Ali Khan, Sher Badshah, Peng Liang, Bilal Khan, Muhammad Waseem, Mahmood Niazi, and Muhammad Azeem Akbar. 2021. Ethics of AI: A Systematic Literature Review of Principles and Challenges. *CoRR* abs/2109.07906 (2021), 1–17. arXiv:2109.07906
[81] Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *JAIR* 70 (May 2021), 871–890. https://doi.org/10.1613/jair.1.12481
[82] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for performing Systematic Literature Reviews in Software Engineering.* Technical Report. Keele University and Durham University Joint Report. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf
[83] Shailendra Kumar and Sanghamitra Choudhury. 2022. Normative Ethics, Human Rights, and Artificial Intelligence. *AI & Ethics* 2 (May 2022), 1–10. https://doi.org/10.1007/s43681-022-00170-8
[84] Alexander Lam, Haris Aziz, Bo Li, Fahimeh Ramezani, and Toby Walsh. 2024. Proportional Fairness in Obnoxious Facility Location. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).* IFAAMAS, Auckland, New Zealand, 1075–1083. https://doi.org/10.5555/3635637.3662963
[85] Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES).* ACM, New York, 86–92. https://doi.org/10.1145/3375627.3375808
[86] Theodore M. Lechterman. 2022. The Concept of Accountability in AI Ethics and Governance. In *The Oxford Handbook of AI Governance.* Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780197579329.013.10 _eprint: https://academic.oup.com/book/0/chapter/386768252/chapter-ag-pdf/50929502/book_41989_section_386768252.ag.pdf.
[87] Michelle Lee, Luciano Floridi, and Jat Singh. 2021. Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics* 1, 1 (June 2021), 529–544. https://doi.org/10.1007/s43681-021-00067-y
[88] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite López-Sánchez, and Juan Rodríguez-Aguilar. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems Through lp-Regression. In *Proceedings of the 21st International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).* IFAAMAS, Virtual Event, New Zealand, 780–788.
[89] Raynaldio Limarga, Maurice Pagnucco, Yang Song, and Abhaya Nayak. 2020. Non-monotonic Reasoning for Machine Ethics with Situation Calculus. In *AI 2020: Advances in Artificial Intelligence.* Springer International Publishing, Canberra, 203–215. https://doi.org/10.1007/978-3-030-64984-5_16
[90] Felix Lindner, Robert Mattmüller, and Bernhard Nebel. 2019. Moral Permissibility of Action Plans. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)* 33, 01 (July 2019), 7635–7642. https://doi.org/10.1609/aaai.v33i01.33017635
[91] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).* IFAAMAS, Virtual Event, London, 799–808. https://doi.org/10.5555/3463952.3464048
[92] Bruce M. McLaren. 2003. Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence* 150, 1 (2003), 145–181. https://doi.org/10.1016/S0004-3702(03)00135-8 AI and Law.
[93] John S. Mill. 1863. *Utilitarianism.* Longmans, Green and Company.
[94] Luís Moniz Pereira and Ari Saptawijaya. 2007. Modelling Morality with Prospective Logic. *International Journal of Reasoning-based Intelligent Systems* 1 (2007), 1–13. https://doi.org/10.1504/IJRIS.2009.028020
[95] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).* IFAAMAS, Virtual Event, London, 907–915.
[96] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. 2019. Norm Emergence in Multiagent Systems: A Viewpoint Paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)* 33, 6 (2019), 706–749.
[97] Luke Munn. 2023. The uselessness of AI ethics. *AI and Ethics* 3, 3 (Aug. 2023), 869–877. https://doi.org/10.1007/s43681-022-00209-w
[98] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent*

Systems (AAMAS). IFAAMAS, Auckland, 1706–1710. https://doi.org/10.5555/3398761.3398958 Blue Sky Ideas Track.

[99] Pradeep K. Murukannaiah and Munindar P. Singh. 2020. From Machine Ethics to Internet Ethics: Broadening the Horizon. *IEEE Internet Computing* 24, 3 (May 2020), 51–57. https://doi.org/10.1109/MIC.2020.2989935

[100] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. 2021. Ethically Compliant Planning within Moral Communities. In *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, Virtual Event, 188–198. https://doi.org/10.1145/3461702.3462522

[101] Samer B. Nashed, Justin Svegliato, and Su Lin Blodgett. 2023. Fairness and Sequential Decision Making: Limits, Lessons, and Opportunities. *ArXiv* abs/2301.05753 (2023), 1–15. https://api.semanticscholar.org/CorpusID:255942265

[102] R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, K. R. Varshney, M. Campbell, M. Singh, and F. Rossi. 2019. Teaching AI agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development* 63, 4/5 (2019), 2:1–2:9. https://doi.org/10.1147/JRD.2019.2940428

[103] Nardine Osman and Mark d'Inverno. 2024. A Computational Framework of Human Values. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Auckland, New Zealand, 1531–1539. https://doi.org/10.5555/3635637.3663013

[104] Ugo Pagallo. 2016. Even angels need the rules: AI, roboethics, and the law. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)* (The Hague, The Netherlands). IOS Press, NLD, 209–215. https://doi.org/10.3233/978-1-61499-672-9-209

[105] Deval Patel, Arindam Khan, and Anand Louis. 2020. Group Fairness for Knapsack Problems. *CoRR* abs/2006.07832 (June 2020), 1–36. arXiv:2006.07832 https://arxiv.org/abs/2006.07832

[106] Erik Persson and Maria Hedlund. 2022. The future of AI in our hands? To what extent are we as individuals morally responsible for guiding the development of AI in a desirable direction? *AI and Ethics* 2, 4 (Nov. 2022), 683–695. https://doi.org/10.1007/s43681-021-00125-5

[107] Michael Pflanzer, Zachary Traylor, Joseph B. Lyons, Veljko Dubljević, and Chang S. Nam. 2023. Ethics in human-AI teaming: principles and perspectives. *AI and Ethics* 3, 3 (Aug. 2023), 917–935. https://doi.org/10.1007/s43681-022-00214-z

[108] Jeremy Pitt. 2022. Contributive Justice and Self-Actualizing Systems. *IEEE Technology and Society Magazine* 41, 4 (2022), 4–11. https://doi.org/10.1109/MTS.2022.3220803

[109] John Rawls. 1967. Distributive Justice. *Philosophy, Politics and Society* 1 (1967), 58–82.

[110] Russell Robbins and William Wallace. 2007. Decision support for ethical problem solving: A multi-agent approach. *Decision Support Systems* 43, 4 (2007), 1571–1587. https://doi.org/10.1016/j.dss.2006.03.003 Special Issue Clusters.

[111] Pamela Robinson. 2023. Moral disagreement and artificial intelligence. *AI and Society* 38, 3 (June 2023), 1–14. https://doi.org/10.1007/s00146-023-01697-y

[112] Manel Rodriguez-Soto, Marc Serramia, Maite López-Sánchez, and Juan Antonio Rodríguez-Aguilar. 2022. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology* 24, 1 (Jan. 2022), 9. https://doi.org/10.1007/s10676-022-09635-0

[113] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating Ethics within Machine Learning Courses. *ACM Transactions on Computing Education* 19, 4 (Aug. 2019), 1–26. https://doi.org/10.1145/3341164

[114] Maureen Sander-Staudt. 2024. Care Ethics. https://iep.utm.edu/care-ethics/. Accessed: 2024-02-17.

[115] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919.

[116] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael Wooldridge, Javier Morales, and Carlos Ansótegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, Stockholm, 1294–1302. https://doi.org/10.5555/3237383.3237891

[117] Marc Serramia, Manel Rodriguez-Soto, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. 2023. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines* 33 (Nov. 2023), 1–30. Issue 3. https://doi.org/10.1007/s11023-023-09649-7

[118] Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* 1, 10 (Oct. 2017), 694–696. https://doi.org/10.1038/s41562-017-0202-6

[119] Zijing Shi, Meng Fang, Yunqiu Xu, Ling Chen, and Yali Du. 2023. Stay Moral and Explore: Learn to Behave Morally in Text-based Games. In *The Eleventh International Conference on Learning Representations (ICLR)*. ICLR, Kigali, Rwanda, 1–18. https://openreview.net/forum?id=CtS2Rs_aYk

[120] Munindar P. Singh. 2013. Norms As a Basis for Governing Sociotechnical Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1, Article 21 (Dec. 2013), 23 pages.

[121] Walter Sinnott-Armstrong. 2021. Consequentialism. In *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford.

[122] Ciske Smit and Jeremy Pitt. 2023. Digital Polycentricity and Value-Sensitive Operationalization. *IEEE Technology and Society Magazine* 42, 4 (2023), 107–118. https://doi.org/10.1109/MTS.2023.3341464
[123] Mohammad Divband Soorati, Enrico Gerding, Enrico Marchioni, Pavel Naumov, Timothy Norman, Sarvapali Ramchurn, Baharak Rastegari, Adam Sobey, Sebastian Stein, Danesh Tarapore, Vahid Yazdanpanah, and Jie Zhang. 2022. From Intelligent Agents to Trustworthy Human-Centred Multiagent Systems. *AI Communications* 35, 4 (2022), 443–457. https://eprints.soton.ac.uk/467975/
[124] Georg Spielthenner. 2005. Consequentialism or deontology? *Philosophia* 33, 1 (Dec. 2005), 217–235. https://doi.org/10.1007/BF02652653
[125] Bernd Carsten Stahl. 2021. *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Springer Cham, New York. https://doi.org/10.1007/978-3-030-69978-9
[126] Jussi Suikkanen. 2017. Consequentialism, Constraints, and Good-Relative-To: A Reply to Mark Schroeder. *Journal of Ethics and Social Philosophy* 3, 1 (2017), 1–9. https://doi.org/10.26556/jesp.v3i1.124.
[127] Ankang Sun, Bo Chen, and Xuan Vinh Doan. 2021. Connections between Fairness Criteria and Efficiency for Allocating Indivisible Chores. *CoRR* abs/2101.07435 (Jan. 2021), 1–32. arXiv:2101.07435 https://arxiv.org/abs/2101.07435
[128] Cass R. Sunstein. 2020. Maximin Special Issue: Regulating the Technological Frontier. *Yale Journal on Regulation* 37 (2020), 940.
[129] Justin Svegliato, Samer B. Nashed, and Shlomo Zilberstein. 2021. Ethically Compliant Sequential Decision Making. *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)* 35, 13 (May 2021), 11657–11665. https://doi.org/10.1609/aaai.v35i13.17386
[130] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2021. Implementations in Machine Ethics: A Survey. *CSUR* 53, 6, Article 132 (Dec. 2021), 38 pages. https://doi.org/10.1145/3419633
[131] Loïs Vanhée and Melania Borit. 2022. Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence. *JAIR* 73 (2022), 619–631.
[132] Carolina Villegas. 2022. *Ethics of Care as Moral Grounding for AI.* Auerbach Publications, Florida, 78–83. https://doi.org/10.1201/9781003278290-13
[133] Carolina Villegas-Galaviz and Kirsten Martin. 2023. Moral distance, AI, and the ethics of care. *AI and Society* 38, 3 (March 2023), 1–12. https://doi.org/10.1007/s00146-023-01642-z
[134] Wendell Wallach, Colin Allen, and Iva Smit. 2008. Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties. *AI and Society* 22, 4 (2008), 565–582. https://doi.org/10.1007/s00146-007-0099-0
[135] Wendell Wallach and Shannon Vallor. 2020. *Moral machines: From value alignment to embodied virtue.* Oxford University Press, United Kingdom, 383–412. https://doi.org/10.1093/oso/9780190905033.003.0014
[136] Lindsay Weinberg. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *JAIR* 74 (2022), 75–109.
[137] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, Honolulu, 195–200. https://doi.org/10.1145/3306618.3314289
[138] Jessica Woodgate and Nirav Ajmeri. 2022. Macro Ethics for Governing Equitable Sociotechnical Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Online, 1824–1828. https://doi.org/10.5555/3535850.3536118 Blue Sky Ideas Track.
[139] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, Timothy J. Norman, and Sarvapali D. Ramchurn. 2022. Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI and Society* 1 (Nov. 2022), 1–12. https://doi.org/10.1007/s00146-022-01607-8
[140] Gary Chan Kok Yew. 2021. Trust in and Ethical Design of Carebots: The Case for Ethics of Care. *International Journal of Social Robotics* 13, 4 (July 2021), 629–645. https://doi.org/10.1007/s12369-020-00653-w
[141] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building Ethics into Artificial Intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI*. IJCAI, Stockholm, 5527–5533.
[142] Jianlong Zhou and Fang Chen. 2023. AI ethics: from principles to practice. *AI and Society* 38, 6 (Dec. 2023), 2693–2703. https://doi.org/10.1007/s00146-022-01602-z

A METHODOLOGY

A.1 Sources Selection and Strategy

After defining our objective and questions, we formed the strategy to search for primary studies by identifying keywords and resources. We selected the University of Bristol Online Library as the resource to search, with Google Scholar as back up. They are both large databases with links

to a wide variety of other sources of research with published papers on the topic. We searched the selected resources using various combinations of the chosen keywords, which can be found in Appendix A.1.1.

Using a forwards and backwards snowballing technique, we inspected up to the first 5 pages of results in each resource, and then narrowed the search by applying the inclusion and exclusion criteria to the titles. This specified the search to a smaller selection of works of whose abstracts were read. The inclusion and exclusion criteria were then more closely applied, identifying primary studies. From the works gathered in this initial search, relevant citations were followed to expand the search, which allowed material to be collected from a broader array of origins. The identification of new key words from the findings was used to update the search string, repeating the process until no new key words were identified.

Figure 2 outlines our search strategy in brief.

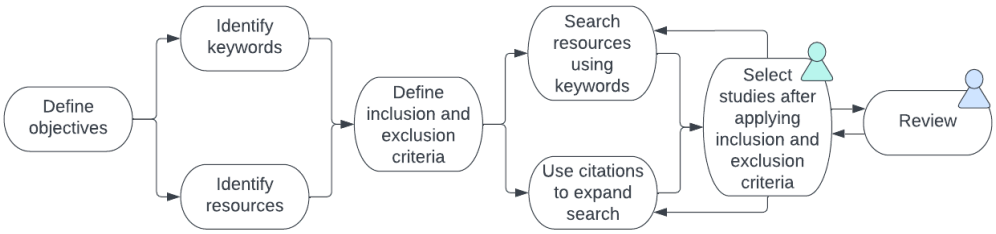

Fig. 2. Search strategy in brief.

A.1.1 Search String Definition. Our search string contained two main components. The first component relates to AI and various related terms, whereas the second component relates to normative ethics. The search string used was ('AI' OR 'Agent' OR 'ML' OR 'Multi-agent' OR 'Multiple-User') AND ('Responsible' OR 'Ethics' OR 'Consequentialism' OR 'Deontology' OR 'Virtue' OR 'Egalitarianism' OR 'Proportionalism' OR 'Kant' OR 'Utilitarianism' OR 'Maximin' OR 'Envy-Freeness' OR 'Doctrine of Double Effect' OR 'Do No Harm').

A.1.2 Inclusion and Exclusion Criteria. First, work is included from a series of well-known journals and conferences identified from literature found in the initial searches. Specifically including these resources ensures topical works are included, however, it also opens up the threat that resources not on the list may be missed. We mitigate risk by following relevant citations from primary studies to expand the scope, however, acknowledge that limitations remain. We exclude works about meta-ethics (e.g., the meaning of moral judgement) and applied ethics outside of AI and computer science (e.g., biology ethics).

Second, we include works about responsible AI. Third, we include works related to individual or group fairness. We exclude works about fairness in specific ML methodology, as that is outside the scope of this project. Fourth, we include the intersection of normative ethics and multiple-user AI research, whereas we exclude studies that do not consider ethics (e.g., studies about technical implementation). Fifth, we include studies about normative ethical principles and AI, but we exclude studies solely about AI principles. This is because this review relates to ethical principles. Sixth, we include studies about bias when related to ethical principles, as this is relevant to how ethical principles affect fairness, however, we exclude studies about bias that do not talk about ethical principles.

Table 9. Inclusion and Exclusion Criteria.

Inclusion
Published works from: ACM CSUR, AIES, FAccT, AAAI, IJCAI, (J)AAMAS, TAAS, TIST, JAIR, AIJ, Nature, Science
Responsible AI
Individual and/or group fairness
Normative ethics and multiple-user AI
Normative ethics and STS
Normative ethical principles and AI
Bias when related to ethical principles

Exclusion
Meta-ethics or applied ethics outside of AI and computer science
Specific ML fairness methodology
Multiple-user AI without reference to ethics
STS without reference to ethics
AI principles without reference to ethical principles
Bias without reference to ethical principles

A.2 Method for Principle Identification

Figure 3 visualises the method used to answer the research questions. This was in a concurrent two-part process of analysing principle identification (Q_p) and principle implementation (Q_o) in literature. Qualitative analysis of works was conducted by reading through and summarising key points, which were then put into relevant classifications of which principles they related to, and their type of contribution (seen in Tables 1 and 2). Classification by principle was conducted by matching papers to the ethical principles which are explicitly stated. Classification by contribution was conducted by utilising categories proposed by Tolmeijer et al. [130] and Yu et al. [141]. These individual analyses were then aggregated to examine the findings as a whole. Some works were more theoretical, exploring the existence of principles and how they might relate to AI and computer science (e.g., Boddington [18]). These works were useful for the identification of principles (Q_p). Other research took established principles and implemented them, which helped to answer Q_o (e.g., Sun et al. [127]). Some works had a mixture of both identification and implementation (e.g., Kim et al. [81]). The first author categorised findings according to principles explicitly stated and contribution as defined above. This analysis was performed in consultation with a second author who critically examined the works being reviewed and the findings extracted by the first author.

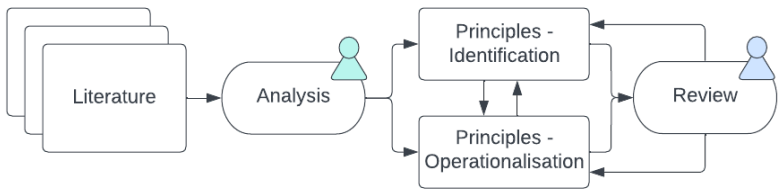

Fig. 3. Methodology to extract principle identification and operationalisation from literature.

A.3 Threats to Validity and Mitigation

Five threats to validity arise, which are summarised here, alongside attempted mitigations. The first threat identified is that only papers that are written or translated into English are included in our

review for developing a taxonomy. This means that relevant research in other languages may be missed, which could contribute to cultural bias and thus threaten both internal and external validity of the study. Internal validity is threatened by missing ethical principles that are referenced in other languages, and the external validity is threatened by diminishing the cross-cultural application of the findings. This is mitigated by seeking papers with an international authorship, but it is recognised as an outstanding issue that could be resolved through future research in applying the methodology to other languages.

A second threat to internal validity is the potentiality of missed keywords, which may again lead to relevant research being excluded. To address this concern, we carefully scope the aims of the review for easier identification of a good array of initial relevant terms. The initial search string is based on preliminary research; as the review continues, more key terms (i.e., ethical principles) are identified. As more terms are identified, a forwards and backwards snowballing technique is used, following relevant citations, updating the search string with new keywords, and repeating the process until no new keywords are identified.

There is a related third threat of missing resources which has similar implications to the internal validity of the study. The topic studied here relates to a broad area of research, and areas such as human-computer interaction and software engineering are not explicitly included in searches but may contain relevant research. This threat is addressed by using two large online libraries as the initial resources, which link to a variety of other resources. Citations from selected studies are also followed, broadening the scope of publications. However, future research could also include reproducing the methodology in these other areas.

Fourth, time limitations threaten the internal validity as there is only time to search the first five pages of results (plus citations). This may mean that there is relevant work beyond these pages that there is not enough time to pursue. To do the best research possible within this time limit, citations are pursued, and Kitchenham and Charters [82] guidelines for a systematic literature review are broadly followed. This helps to effectively identify relevant research. On the other hand, this limitation could lead to further research in this area by applying our methodology to the analysis of more studies than those identified here.

The fifth issue of researcher bias also threatens internal validity as it can sway the results in a particular direction rather than being objective. This is mitigated by having a secondary reviewer who critically analyses results and makes suggestions to help the primary reviewer improve the study. This is also tackled by basing the study selection criteria on the research question and defining it before the review is begun.