Normative 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 decisions that consider human values and can be justified by human morals. Operationalising normative ethical principles inferred from philosophy supports responsible reasoning. We survey computer science literature and develop a taxonomy of 23 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 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 moral responsibility [26, 89]. These considerations are captured under the perspective of sociotechnical systems (STS), which incorporates the human element in ethical reasoning [79]. Within the concept of STS, it is also important to adopt the perspective of macro ethics [21, 104], which considers the role of values (what is important to us in life [91]) and other ethical features, rather than examining ethics from the atomic perspective of a single agent [81, 90]. This is because ethics should be understood as a reflective development process that incorporates context [62, 71, 76, 107]. Values are an important aspect of context [70] as they reflect stakeholder preferences [35].

To ensure AI systems behaves in responsible ways, reasoning capacities of these systems should thus consider relevant values, morals, and ethical considerations [32, 95, 105]. However, users (here, users are synonymous with stakeholders) may have different value preferences, or their values may conflict with norms (rules of expected behaviour [77]) [27]. We argue that incorporating normative ethical principles is a step forward to creating responsible AI systems.

1.1 Motivation for a Taxonomy of Ethical Principles

The motivation for this work stems from the need to improve ethical considerations in reasoning capacities for responsible AI. Operationalising principles from normative ethics enables systems to methodically reason about ethics [104]. 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 [79]. Ethical principles imply certain logical propositions that must be true for a given action plan to be ethical [60]. The application of ethical principles is useful to methodically think through dilemmas and promote satisfactory outcomes [24]. These principles guide normative judgements, determine the moral permissibility of concrete courses of action and help to understand different perspectives [17, 69, 73, 88].

Normative ethical principles have previously been utilised for a variety of different applications in computer science. Works such as Binns [11] and Leben [64] apply ethical principles to improve fairness considerations for binary machine learning algorithms. Cointe et al. [23] implement ethical principles in decision making, enabling agents to make ethical judgements in specific contexts.

Authors' address: Jessica Woodgate, yp19484@bristol.ac.uk; Nirav Ajmeri, nirav.ajmeri@bristol.ac.uk, University of Bristol, Bristol, United Kingdom, BS8 1UB.

Ethical principles can also be applied to improve fairness considerations in systematic analysis [24, 88].

Ethical thinking should be fostered through the appreciation of a variety of different approaches, considering the strengths and limitations of each [15, 85]. We envision that a taxonomy of ethical principles, including how they have previously been operationalised, will contribute to the overall aim of improving fairness in the reasoning of responsible AI.

1.2 Gaps in Related Research

In the context of AI ethics, there are two types of principles referred to (1) those inferred from normative ethics such as Deontology and Consequentialism, as found in Leben [64], and (2) those adapted from other disciplines like medicine and bioethics such as those suggested by Floridi and Cowls [43], Jobin et al. [54], Fjeld et al. [41], Whittlestone et al. [103], Cheng et al. [20], and Bo et al. [66] including beneficence, non-maleficence, autonomy, justice, fairness, non-discrimination, transparency, responsibility, privacy, 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 [43] and Jobin et al. [54] as *AI principles*. Ethical principles are thus philosophical theories which can be operationalised in reasoning capacities, as they imply certain logical propositions which must be true for a given action plan to be ethical [60]. These principles broadly divide into Deontological principles (those which entail conforming to rules, laws, and norms [49]) and Teleological principles (those which derive duty or moral obligation from what is good or desirable as an end to be achieved [14]). AI principles relate to what kinds of things ought to be promoted in the development and use of AI.

AI Principles Themes such as beneficence, non-maleficence, autonomy, justice, fairness, non-discrimination, transparency, responsibility, privacy, accountability, safety and security, explainability, human control of technology, and promotion of human values that should underpin the design of AI technologies.

Ethical Principles Operationalisable rules inferred from philosophical theories such as Deontology and Consequentialism.

Existing taxonomies and surveys are present in the relevant but distinct domain of AI principles such as Jobin et al. [54], Floridi and Cowls [43] and Khan et al. [59], however, they consider AI principles rather than ethical principles. Dignum [33], Leben [64], and Robbins and Wallace [85] give summaries of normative ethics. The work of Tolmeijer et al. [100] contains 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. [106] provide a concise guide to ethical dilemmas in AI and identify a high-level overview of ethical principles.

1.3 Novelty

We build upon previous research, especially that of Tolmeijer et al. [100], to collate a broader range of ethical principles discussed in the computer science literature, summarising operationalisation, principle by principle. There are two key aspects of novelty contributed by this paper.

Broadening the Range of Ethical Principles We identify a taxonomy tree with 23 ethical principles discussed in computer science literature.

Principle Specific Operationalisation We identify 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.

1.4 Organisation

Section 2 explains our methodology in brief. This may be useful for future research seeking to expand the taxonomy of ethical principles by reproducing the methods used here. Section 3 explores our findings from our objective to identify ethical principles that have been so far proposed in computer science literature. Section 4 examines our objective to look at how ethical principles have been operationalised in previous literature, and what steps practitioners seeking to operationalise principles should take. Section 5 focuses our objective to identify gaps in operationalising ethical principles in computer science and AI, and future directions these gaps may lead to. Section 6 concludes with our key takeaways.

2 METHODOLOGY

Taking inspiration from software engineering research, for reproducibility we follow Kitchenham's [61] 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. An initial search string is constructed from preliminary research. Using a forwards and backwards snowballing technique, we search our selected resources (the University of Bristol library, with Google Scholar as backup) using our search string. After applying our inclusion and exclusion criteria to identify our primary studies, relevant citations are followed to expand the search. The identification of new key words is used to update the search string, repeating the process until no new key words emerge. For further details of the methodology, see Appendix A.

2.1 Objective

Our broad objective is to 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 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 the approach of Tolmeijer et al. [100] in providing brief overviews of how each principle has been defined in literature, rather than attempting to give an introduction to moral philosophy.

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 [64] and Tolmeijer et al. [100] offer guidance as to how certain 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 23-May-22. The 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 lead to around 10–20 studies from

Table 1. Categorisation of Research Reviewed with Principles Extracted: Frameworks.

	Frameworks (Conceptualization)	Frameworks (Application)	Algorithms	Viewpoint or Review
Deontology	[1, 11, 13, 23, 48, 64, 79, 88, 102]	[5, 8, 10, 28, 38, 52, 67, 69, 85]	[52, 86]	[9, 49, 58, 96, 100, 106]
Egalitarianism	[8, 11, 22, 37, 42, 44, 64, 78, 84, 92]	[36]	–	[65]
Proportionalism	[39, 56, 64]	[36]	–	–
Kantian	[1, 3, 50, 57, 60, 102]	[10, 38, 67, 85, 99]	[86]	[63, 100]
Virtue	[1, 4, 13, 23, 48, 50, 52, 79, 88, 102]	[47, 52, 85]	[52, 86]	[9, 49, 58, 100, 101, 106]
Consequentialism	[1, 13, 23, 25, 48, 50, 64, 88, 94, 97]	[8, 10, 67]	[86]	[9, 40, 96, 100, 106]
Utilitarianism	[1, 3, 4, 8, 13, 52, 60, 64, 74, 78, 79, 102]	[2, 6, 10, 18, 28, 67, 69, 85, 99]	[19, 52, 86, 93]	[40, 58, 63, 96, 106]
Maximin	[64, 83]	[2, 10]	[19, 31, 98]	[65]
Envy-Freeness	[12]	–	[98]	[65]
Doctrine of Double Effect	–	[10, 46, 69, 75]	–	[29]
Doctrine of Disparate Impact	[11]	–	[82]	–
Do No Harm	[30]	[69]	–	–

each resource. Closer examination of these works lead to the identification of relevant citations and which we incorporated into our review. The selection of these works was critiqued by a secondary researcher which helped the identification of further relevant research. This resulted in 57 papers being included in the review. We conducted a second search on 14-January-23, resulting in a further 10 papers being included in the review.

3 TAXONOMY OF ETHICAL PRINCIPLES

We now address Q_p (Principles) on identifying ethical principles proposed so far. Research on AI and ethical principles is broadly categorised into twelve key principles, based on their definition of normative ethics (Deontology, Egalitarianism, Proportionalism, Kantian, Virtue, Consequentialism, Utilitarianism, Maximin, Envy-Freeness, Doctrine of Double Effect, Doctrine of Disparate Impact, and Do No Harm), and five types of study (Frameworks (Conceptualisation), Frameworks (Application), Algorithms, and Viewpoint or Review), based on the structure and contributions of the paper. Table 1 shows this broad categorisation.

In terms of paper structure, we find that a large majority of works focus on conceptual frameworks which propose theoretical ideas as to how the principles might be operationalised, such as Leben [64] and Wallach et al. [102]. Some papers computationally apply such frameworks, for example, Limarga et al. [67] and Linder et al. [69]. A few papers propose algorithms mechanising ethical principles, such as Sun et al. [98] and Diana et al. [31]. Lastly, viewpoint or review papers scope related areas, such as Yu et al. [106] and Lee et al. [65].

Within normative ethical principles, there are two main strands of theory: *Deontology* and *Teleology*. Deontological theories revolve around rules, rights, and duties [79, 102]. Teleological ethics, on the other hand, derive duty or moral obligation from what is good or desirable as an end to be achieved, [14]. Teleological ethics can be further divided into Consequentialism, Egoism, and Virtue Ethics. Figure 1 displays the taxonomy of principles identified in literature in a tree structure, mapping out how they relate to each other.

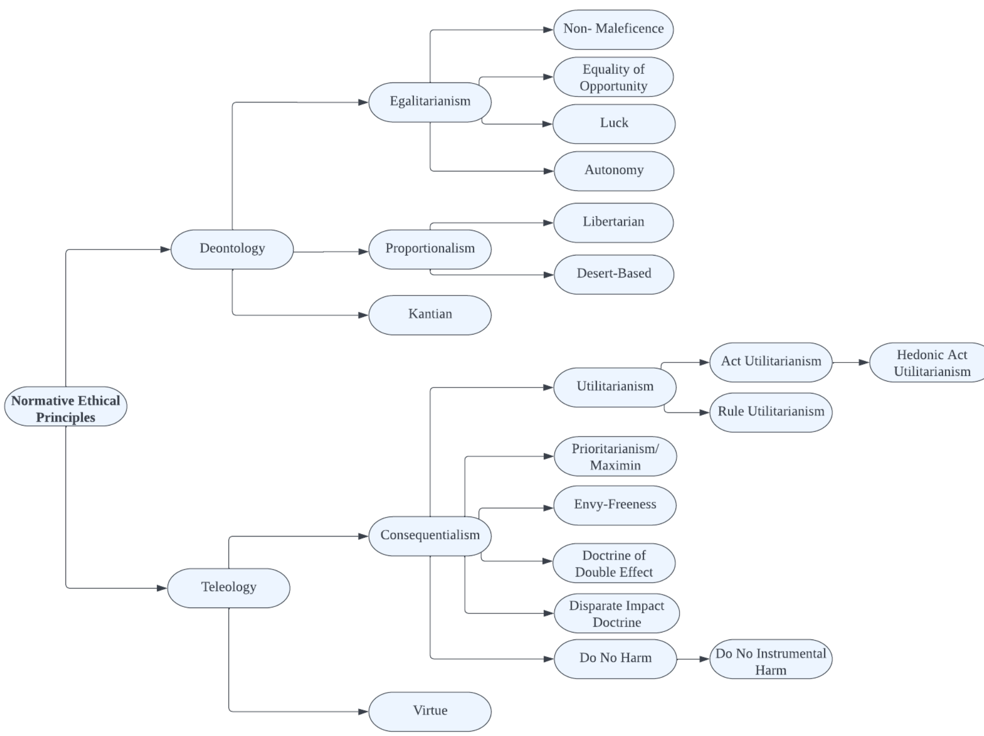

Fig. 1. Taxonomy of Ethical Principles.

We find that certain principles, such as Utilitarianism, are much more discussed than other principles such as Do No Harm, as can be seen in Table 1. 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, Cointe et al. [23], Greene et al. [48], and Anderson and Anderson [5]. These works would perhaps benefit by more clearly specifying the ethical principles they are using, to allow for more precise operationalisation.

3.1 Deontology

Deontology entails conforming to rules, laws, and norms [49, 79], and respecting relevant obligations and permissions [23] that stem from duties and rights [3, 58, 85, 86, 88, 100, 102, 106]. For Deontological theories, the permissibility of action lies within the intrinsic character of the act itself. An action is permissible if and only if the act itself is intrinsically morally good, independent of the outcome [13, 67, 69].

To implement Deontological theories, a rules-based approach may be used to identify appropriate actions: Limarga et al. [67] use predicates to encode rules, and then reason about different types of actions; Berreby et al. [10] first collect contextual information to simulate the outcome of actions, and then assess the ethical considerations of that outcome using Deontological specifications; Tolmeijer et al. [100] argue that Deontology could be implemented by inputting the action (in terms of mental states and consequences), 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 [11] uses Deontology to choose between incompatible fairness metrics, whereas Leben [64] applies it to evaluate distributions of binary classification algorithms. Some works also suggest using Deontology only in specific circumstances: Dehghani et al. [28] choose to implement Deontology in situations with 'sacred values', using it to select the choice of action that doesn't violate the sacred value.

However, there are issues that 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. This makes it challenging for logic to adequately capture complex ethical insights [1, 88]. Also, rights-based ethics revolve around decisions based on the rights of those who are affected by the decision, but this can be less helpful in situations where rights are not impinged yet some sort of ethical dilemma is still occurring. In terms of implementation, there may be issues that arise when exceptions to the rule emerge. Rules are expected to be strictly followed, implying that for every exception they must be amended. This could result in very long rules. Determining the right level of detail is thus important to ensure interpretability for the machine [100]. 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.

3.1.1 Egalitarianism. Egalitarianism stems from the notion that human beings are in some fundamental sense equal. To administer Egalitarianism, efforts should be made to avoid and correct certain forms of inequality [11].

Literature implements Egalitarianism by promoting equality in different ways: Murukannaiah et al. [78] suggest minimising disparity across stakeholders with respect to satisfying their preferences; Dwork et al. [36] classify individuals who are similar with respect to a particular attribute similarly; for resource allocation, Leben [64] confers equal rights (and thus equal shares) to each member of the population. However, if it is impossible to achieve equality across all metrics for the entire population, they suggest a distribution that minimises the distance to some fairness standard (e.g., size of population).

Egalitarianism may be applied to evaluate various algorithmic fairness metrics, such as positive predictive parity or equal odds, Lee et al. [65] proposes. This helps developers to decide what layers of inequality should (not) be influencing a model's prediction.

Certain difficulties are important to acknowledge when considering Egalitarianism. For instance, there is 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 [11]. 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 lower socioeconomic backgrounds being excluded from democracy. However, having 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 2 describes sub-types of Egalitarianism.

Table 2. Subtypes of Egalitarianism.

Principle	Description	Difficulties
Non-Maleficence	Imposes Egalitarianism across harms but not benefits [64]. In optimisation techniques, different actions could be assigned values based on a predetermined formula, identifying the 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' [64]. 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, yet individuals should be still held accountable for their actions [36, 44]. Opportunities should therefore be equally distributed. One could examine whether each group is equally likely to be predicted a desirable outcome given the base rates for that group [11], or ensure all opportunities should be equally open to all applicants based on a relevant definition of merit [65].	In theory, this can be fully satisfied even if only a minority segment of the population has realistic prospects of accessing the opportunity [42].
Luck	Inequalities that stem from unchosen aspects should be eliminated so that no-one is worse off due to bad luck. Instead, people should be given differentiated benefits as a result of their own choices [37, 65]. From an optimisation perspective, stakeholders could be given a weighting which mitigates the effects of luck. Allocations are then distributed equally, accounting for this weighting.	It is often difficult to define what is within an individual's genuine control [11]. The ideal solution would allow inequalities resulting from people's free choices and informed risk-taking, disregarding those which are the result of luck.
Autonomy	Levels of autonomy should be equally distributed, through a variety and quality of options, and decision-making competence [42]. The aim of this would be to incorporate the full range of individual freedom [65]. Levels of autonomy could be inputted to reasoning about potential actions, selecting the action with the most equal distribution of autonomy.	However, when there is a significant asymmetry of power and information, autonomy in rational decision-makers fails as an ethical objective [42].

3.1.2 Proportionalism. Proportionalism infers adjusting the rights of each person proportionally based on their contributions to production. Depending on the sub-type of proportionalism (shown in Table 3), 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 luck that went into those resources. Leben [64] operationalises Proportionalism by constructing utility functions which 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 the amount of each individual's contribution. The best distribution is that which has the minimum distance from this fairness standard for all individuals.

A challenge with Proportionalism is that there may be situations in which groups or individuals did not confer contributions to production, but should still be granted a distribution of rights. For example, those unable to contribute due to disability should still have a fair distribution of rights. This may be mitigated by considering the influence of luck. For sub-types of Proportionalism, see Table 3.

Table 3. Subtypes of Proportionalism.

Principle	Description	Difficulties
Libertarian	Libertarianism emphasises the importance of each person's freedom, insofar as there is no harm to anyone else [65]. Rights are distributed accordingly with each person's total contribution at the time of consent. Each group is entitled to success rates at least as fair as initial contributions [64].	A difficulty with this approach lies in that it does not target pre-existing inequalities which may still be worth mitigating. For example, some people may be inhibited from contributing as much as others due to factors outside of their control such as generational wealth inequality. On the other hand, people born into wealthy circumstances due to luck would be rewarded more rights according to this approach, which may seem unfair.
Desert-Based	Contribution is defined in terms of individual effort, discounting the effects of luck, and rights are distributed accordingly. This is because the prior prevalence of a trait in a population can be the result of unjust circumstances [64]. This could be implemented by assigning each individual some distance in a metric space that evaluates desert, and evaluating the fairness through the average distance between individuals from each group in the metric space [36].	A weakness of this principle is that luck is an abstract concept which is difficult to define, and may vary between contexts. It is thus challenging to evaluate which traits should be mitigated for.

3.1.3 Kantian. Kantian [57] ethics argues that ethical principles are derived from the logical structure of action, beginning with distinguishing free action (behaviour for which the agent has reasons) from mere behaviour [60]. Kant's *Categorical Imperative* grounds all moral duties [102], as it applies unconditionally to rational agents (categorical), and is a command that could be followed, but might not be (imperative) [55]. 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*) [1, 63, 85]. 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 can kill but should not. Derived from the Categorical Imperative is the *Means-End Principle* (also known as the *Humanity Formula*). This denotes that treating other people as a means to an end is immoral [1, 63]. 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 is because of our ability to engage in rational self-directed behaviour.

This principle has been operationalised in previous literature through the imposition of rules. Limarga et al. [67] 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. [10] 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. [99] decouple moral principles (in the form of rules) from the decision-making module; for Kantianism, they use the moral rule that policies should be universalisable to stakeholders without contradiction. Allen et al. [3] 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; 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 universalizable maxim [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.

3.2 Teleological Ethical Principles

This section examines each Teleological (morality is derived from what is good or desirable as an end to be achieved [14]) principle identified in the review, including details as to how they have been previously operationalised and difficulties that may arise.

3.2.1 Virtue Ethics. According to Virtue Ethics, ethicality stems from the inherent character of an individual, and not the rightness or wrongness of individual acts [1, 13, 58, 79, 85, 102, 106]. 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 [1, 4, 86, 88]. 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 [102].

Virtue Ethics has been implemented in literature in a variety of ways. Robbins and Wallace [85] argue that to operationalise this principle, problems are solved through the application of 'virtuous' characteristics. This may be done by targeting the designers of systems, and helping designers develop virtues through education, as suggested by Vanhé and Borit [101]. Other works focus on implementing virtues directly into machines; according to Tolmeijer et al. [100], 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. This is exemplified by Govindarajulu et al. [46] who define virtues as learnt by experiencing the emotion of admiration when observing virtuous people, and then copying the traits of those people. The authors instantiate this by using computational formal logic to formalise emotions (in particular the emotion of admiration), represent (virtuous) traits, and establish a process of learning traits. Greene et al. [48] argue that a virtue-based system would have to appreciate the entire variety of features in a given situation that would call for one action rather than another. Virtue Ethics can also be used alongside other approaches; Hagendorff [49] argue that Deontological approaches should be combined with Virtue Ethics, by looking at values and character dispositions.

A problem with Virtue Ethics is that the holistic view it takes makes it more difficult to apply to individual situations [88]. Further challenges relate to conflicting virtues and the concretion of virtues [100]. 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 – the reasons behind the actions need to be clear. This therefore makes it difficult to build virtues into machines, as there is a high level of abstraction to what virtues are. Additionally, the conception of virtues can change greatly across time and culture. Virtues instantiated in machines now may lead to unfair outcomes in the future as virtues change.

3.2.2 Consequentialism. In Consequentialist approaches, right actions are identified through their effects [13, 23, 64, 106]. The moral validity of an action can thus be judged only by considering its consequences [67, 86, 88]. 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 these 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 [94].

Consequentialist principles can be operationalised by analysing the consequences of different actions. This is different to Deontology, which regards 'mental states' as very important for determining the ethicality of an action. For Consequentialism, mental states can be largely disregarded [100]. Consequentialism is implemented by Limarga et al. [67], who assign each action a weight according to its worst consequence. Actions are part of a sequence of actions to reach a goal, and their weights accumulate to a total amount. This total amount is then optimised to select the sequence with the best overall consequence. Suikkanen [97] 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. [100] 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 then be mechanised by selecting the consequence with maximum utility. For binary classification algorithms, Leben [64] suggests implementing Consequentialism by looking at how weights are assigned to each group outcome based on relative social cost.

In practice, assigning weights to each outcome may be unrealistic to do for all outcomes [64]. There might be high computational costs, because Consequentialist systems would require a machine to represent all of the actions available to it [48]. A related issue is that it is difficult to estimate long-term or uncertain consequences and determine which consequences should be taken into account [40, 88]. There may also be moral constraints outside Consequentialism which prohibit certain actions even when they have the best outcomes, therefore rendering Consequentialist theories incomplete [97]. 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.) [13, 100]. For sub-types of Consequentialism, see Table 4.

Table 4. Subtypes of Consequentialism.

Principle	Description	Difficulties
Utilitarianism	The ultimate end is an existence exempt as far from pain and as rich in enjoyment as possible [3, 74]. Something is ethical if and only if it maximizes the total net expected utility across all who are affected [58, 60, 63, 86]. Agents could thus be trained to make judgements that deliver the greatest happiness to the greatest number of people [63], e.g. by assigning a value to every action which is used for final evaluation [67]; selecting the choice with the highest utility [8, 28]. One approach decouples the moral principle from the decision module, having a separate moral rule that evaluates the suggested policy [99]. Utilitarianism can be applied to choose norms by how much they align with moral values, by ensuring norms that promote the most utility are chosen, where utility is understood as (value) preference satisfaction. A recursive utility function identifies the preference utility of each value; the value support of a norm is then calculated by adding the utility of each value that norm is supported by [93]. To justify design choices for fairness metrics in binary classification algorithms, model each potential distribution and its effects (a utility function/measure of happiness outcomes); then run a selection procedure over aggregate utilities to maximise the sum [64].	This could lead to a minority being treated unfairly for the greater good [2, 6]. It may be impossible to calculate the utility of every action, and the theory cannot account for the notion of rights and duties [1]. It is also difficult to quantify utility [40]. To mitigate these issues, Utilitarianism could be an additional necessary condition for ethical action, rather than the sole ethical principle [60].
(Hedonic) Act Utilitarianism	Morality of action lies in its consequences [100]; Hedonic Act Utilitarianism entails computing the best action which derives the greatest net pleasure [13]. 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 [10]. Alternatively, one could input the number of people affected, and for each person the intensity of pleasure/displeasure 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 [6].	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. This makes it difficult to identify the action with the greatest net pleasure.

Principle	Description	Difficulties
Rule Utilitarianism	Morally assessing an action 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. 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 [10].	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.
Prioritarianism/Maximin	Maximising the minimum utility by seeking to improve the worst-case experience in a society [65, 83], e.g. an agent that aims to improve the minimum experience/worst-case outcome for any user [2]. For choices of fairness metrics for binary classification algorithms, a function modelling each potential distribution and its effects could be constructed, and then a selection procedure run over aggregate utilities [64]. Fairness could be measured by worst-case outcomes across all groups, rather than differences between group outcomes [31], or by minimising the maximum cost of an allocation over all allocations [98]. Maximin could also be coupled with Utilitarianism in optimisation problems, to ensure the least advantaged have priority, but not at unlimited cost to everyone else [19].	Although the aggregate utility may be increased, it does not necessarily mitigate the effects of discrimination [98]. It still allows for disparities between groups. Therefore, the most privileged group may remain much more privileged than the least privileged group, despite the overall experience being improved.
Envy-Freeness	In an Envy-Free allocation, no agent envies another agent [98]. 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 [12].	Arguably, what is important might not be a relative condition to other people, but if people have enough to have satisfactory life prospects [65]. 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 [98].
Doctrine of Double Effect	Deliberately inflicting harm is wrong, even if it leads to good [29, 75]. An action is permissible if the action itself is morally good/neutral, some positive consequence is intended, no negative consequence is a means to the goal, and the positive consequences sufficiently outweigh the negative ones [10, 46, 69].	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.
Doctrine of Disparate Impact	For group fairness, any group must have approximately equal or proportional representation in the solution provided by the algorithm [82]. The concept of 'disparate mistreatment' has also been suggested, which considers differences in false positive rates between groups [11]. It thus emphasises the importance of ensuring impact is proportionally distributed amongst the relevant groups.	This is may lead to individuals being unfairly treated in favour of the group.

Principle	Description	Difficulties
Do No (Instrumental) Harm	People should be free to act as they wish unless doing so would result in harm to another person [30, 45, 69]. Do No Instrumental Harm allows for harm as a side effect, but not as a means to a goal.	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.

3.3 Other Principles

In addition to the principles mapped out here, there are other principles mentioned in literature which we now describe.

3.3.1 Egoism. Egoism is acting to reach the greatest outcome possible for one's self, irrespective of others [63, 85]. 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 act in a way which is aligned to human values.

3.3.2 Particularism. Particularism emphasises that there is no unique source of normative value, nor is there a single, universally applicable procedure for moral assessment [100]. Rules or precedents can guide evaluative practices, however they are deemed too crude to do justice to many individual situations. Therefore, whether a certain feature is morally relevant or not and what role 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. 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.

3.3.3 The Ethic of Care and Responsibility. The ethic of care and responsibility emphasises considering your feelings of interconnectedness with others [72, 85]. To be ethical, one should think about the situation that each of these others and you are in. Using your experience, you should act in a nurturing and responsible way. This is a key guiding factor to have in the application of ethical principles, as it enhances the importance of considering others outside of yourself. This provides good support for value alignment and responsible decision making.

3.3.4 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 [50]. 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. This is crucial to 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 conduct an analysis of previous operationalisation of ethical principles for Q_o (Operationalisation). We find a variety of techniques used for the technical implementation of ethical principles, summarised in Table 5 and Table 6. We find that previous literature integrates principles into reasoning capacities in a top-down, bottom-up, or hybrid architecture, summarised in Table 7. Practitioners should be specific about which principle(s) they are operationalising, and previous literature suggests that Pluralism may help with this decision. We also find that, abstractly, operationalisation falls into the categories of applying rules for Deontological principles, developing virtues for Virtue Ethics, or evaluating consequences for Consequentialist principles.

4.1 Choosing Technical Implementation

In previous literature, a variety of technical implementations have been used to encode ethical principles. Expanding upon Tolmeijer et al.'s categorisation [100], approaches to encode principles into a format that computers can understand include logical reasoning, probabilistic reasoning, learning, optimisation, and case-based reasoning [87]. In Table 5 and Table 6, we map each ethical principle found in literature to their technical implementations.

4.2 Clarifying the Architecture

To engineer morally sensitive systems, practitioners must decide on the architecture for integrating ethical principles [102]. These fall within three broad approaches: top-down imposition of ethical theories; bottom-up building of systems with goals that may or may not be explicitly specified; and hybrid approaches which combine top-down and bottom-up features. Table 7 summarise our findings of how ethical principles have been implemented according to the different 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 [40]. Bottom-up techniques suggested by Tolmeijer et al. [100] include using artificial neural networks, reinforcement learning, and evolutionary computing. An example of this is Noothigattu et al. [80], 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. [80] align policies with human values. This may improve explainability by assimilating policies with principles which, by their nature, imply logical propositions that can be reasoned about [60]. Dyoub et al. [38] utilise answer set programming (ASP) as the main knowledge representation and reasoning language to deductively encode ethical rules. They then utilise inductive logic programming to learn 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 [60], instead of asking the machine to learn from experience (as in bottom-up approaches). Top-down approaches are rule-based: ethics is understood as the investigation of right actions through identifying rules that should be followed to perform the morally correct (or at least permissible) action [68]. We find many works use top-down approaches to integrate ethical principles into the reasoning capacities of machines. Dehghani et al. [28] implement Deontological and Utilitarian principles through a combination of qualitative modelling, first-principles logical reasoning, and

Table 5. Technical Implementation of Principles.

Implementation Type		Ethical Principles				
		Deontology	Egalitarianism	Proportionalism	Kantian	Virtue
Logical Reasoning	Deductive Logic	–	–	–	[85]	[85]
	Non-Monotonic Logic	–	–	–	[10, 67]	–
	Abductive Logic	–	–	–	–	–
	Deontic Logic	[69]	–	–	–	[47]
	Rule-Based Systems	[23, 28]	[8]	[39]	[85]	[23, 85]
	Event Calculus	–	–	–	[10, 67]	[47]
	Knowledge Representation and Ontologies	[23, 28]	–	–	[85]	[23, 85]
	Inductive Logic	[5, 38]	–	–	[38]	–
Probabilistic Reasoning	Bayesian Approaches	–	–	–	–	–
	Markov Models	[86]	–	–	[86, 99]	[86]
	Statistical Inference	–	[36]	[36]	–	–
Learning	Decision Tree	–	[8]	–	–	–
	Reinforcement Learning	[86]	–	–	[86]	[86]
	Inverse Reinforcement Learning	[80]	–	–	–	–
	Neural Networks	[52]	–	–	–	[52, 53]
	Evolutionary Computing	–	–	–	–	[53]
Optimisation		–	[6, 8, 36, 64, 93]	[36, 64]	–	[6]
Case-Based Reasoning		[28, 73]	–	–	–	–

analogical reasoning. Diana et al. [31] operationalise the principle of minimax (minimising the maximum loss, adapted from Maximin - maximise the minimum) using oracle-efficient learning algorithms. They apply minimax to analyse fairness considerations in differences between group outcomes. Also considering fairness, Sun et al. [98] formalise Envy-Freeness as rules to examine the trade-off between different fairness allocations. Chen and Hooker [19] combine the principles of Maximin and Utilitarianism in a model for mixed integer/linear programming, which can be applied in a top-down manner to optimise social welfare functions. Tolmeijer et al. [100] 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. However, as human knowledge does not tend to be very structured, it needs to be interpreted before it can be used. A difficulty of top-down approaches is thus 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 in hybrid approaches may result in better implementation of ethical principles [3]. Hybrid examples include Berreby et al. [10], who 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. [67] 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. [86] suggest 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. Their algorithm builds an environment where it is in the best interest of the agent to act ethically while still pursuing its individual objective. To consider several objectives, they model the environment as a multi-objective Markov decision process, which allows the agent to consider both individual and ethical objectives. 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. A benefit of hybrid approaches is that they incorporate both ethical reasoning and empirical observation, which allows context to be taken into account.

Table 7. Architecture for Implementing Principles.

Ethical Principles	Bottom-Up	Top-Down	Hybrid
Deontology	Inverse Reinforcement Learning [80] Inductive Logic [5]	Rule Based Systems, Knowledge Representation and Ontologies, Case-Based Reasoning [28] Deontic Logic [69] Case Based Reasoning [73]	Neural Networks [52] Rule-Base Systems, Knowledge Representation and Ontologies [23] Markov Models, Reinforcement Learning [86]
Egalitarianism	–	Statistical Inference, Optimisation [36] Optimisation [6, 64, 93]	Rule-Based Systems, Decision Tree, Optimisation [8]

Ethical Principles	Bottom-Up	Top-Down	Hybrid
Proportionalism	–	Statistical Inference, Optimisation [36] Rule-Based Systems [39] Optimisation [64]	–
Kantian	–	Deductive Logic, Rule-Based Systems, Knowledge Representation and Ontologies [85] Markov Models [99]	Non-Monotonic Reasoning and Event Calculus [10, 67] Markov Models, Reinforcement Learning [86]
Virtue	Evolutionary Computing [53]	Deductive Logic, Rule-Based Systems, Knowledge Representation and Ontologies [85]	Neural Networks [52] Deontic Logic, Event Calculus [47] Rule-Base Systems, Knowledge Representation and Ontologies [23] Markov Models, Reinforcement Learning [86]
Consequentialism	–	–	Rule-Base Systems, Knowledge Representation and Ontologies [23] Markov Models, Reinforcement Learning [86]
Utilitarianism	Rule Based Systems, Knowledge Representation and Ontologies [28]	Deductive Logic, Rule-Based Systems, Knowledge Representation and Ontologies [85] Deontic Logic [69] Optimisation [6, 64]	Non-Monotonic Reasoning and Event Calculus [10, 67] Neural Networks [52] Rule-Based Systems, Decision Tree, Optimisation [8] Bayesian Approaches, Optimisation [7]
Maximin	–	Optimisation [31, 64, 82] Rule-Based Systems, Reinforcement Learning [2]	Non-Monotonic Reasoning and Event Calculus [10]
Envy-Freeness	–	Optimisation [98]	
Doctrine of Double Effect	–	Abductive Logic [75] Deontic Logic [69] Deontic Logic, Event Calculus, Markov Models [46]	Non-Monotonic Reasoning and Event Calculus [10]

Ethical Principles	Bottom-Up	Top-Down	Hybrid
Doctrine of Disparate Impact	–	Optimisation [11, 82]	–
Do No Harm	–	Deontic Logic [69] Rule-Based Systems [30]	–

4.3 Specifying the Ethical Principle

Practitioners should specify which ethical principle(s) will be operationalised. This could be aided by the taxonomy we have suggested, which contains a broad array of ethical principles found in AI and computer science literature, with 23 nodes (Figure 1). Being clear about which principle is being used will help designers to further specify what inputs are necessary for their application, which in turn will improve the ethical reasoning capabilities and explainability of how decisions have been made [64].

In specifying the ethical principle, it may be useful to apply the theory of Pluralism. Pluralism states that there is no one approach that is best [85], as human morality is complex and cannot be captured by one single classical ethical theory. Context and various reasoning techniques could be used to choose between appropriate principles. Tolmeijer et al. [100] advocate for further research according to this approach, suggesting the development of multi-theory models where machines can interchangeably apply different theories depending on the type of situation. An example of pluralism being operationalised can be found in Svegliato et al. [99], who propose a framework in which ethical compliance is decoupled from task completion to avoid unanticipated scenarios which do not reflect the values of stakeholders. They suggest that this can be done by implementing a pluralist approach in the form of having an extra moral constraint which represents a moral principle. This allows for the morality of a policy of the decision-making module to be evaluated considering its ethical context, leaving room for different ethical principles to be implemented as the ethical rule. Pluralism is a useful approach to have for ethics, and help AI practitioners decide which ethical principle(s) are appropriate.

4.4 Using Rules, Consequences, or Virtues

Depending on the type of principle being used, previous works have operationalised principles in three main ways. Deontological principles have been operationalised through applying rules, and choosing an action based on how it accords with certain rules. Virtue Ethics has been operationalised through 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], Greene et al. [48], and Berreby et al. [10]. Examples of this would be applying the rule that the disparity of preference satisfaction for stakeholders should be minimised, extracted from the principle of Egalitarianism [78]. Another example is applying the rule that stakeholders should be treated proportionally based on their contributions to production [64]. However, due to the abstract nature of ethics, some 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 [100]. For example, an essential part of Kantianism [57] is that the reasons for actions must be universalizable to all agents and therefore perhaps this rule should be strict. However, arguably this could permit actions that are bad according to other principles [1], suggesting that it should be defeasible. Creating systematic ways of encoding the whole taxonomy of 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 [1, 13, 58, 79, 102, 106]. To solve a problem according to this theory, virtuous characteristics should be applied [85]. Thus, the theory can be operationalised through instantiating virtues [100]. This is exemplified by Govindarajulu et al. [46], 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 instantiated by using computational formal logic to formalise emotions (in particular the emotion of admiration), represent traits (which in this example will be virtuous), 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 (which can take sentences as arguments and allows for possible states [87]) 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 these traits the agents learn the traits, instantiating virtues. However, Virtue Ethics can be difficult to apply to individual situations [88], and there are thus challenges that arise with the application of virtues across time and culture [100]. 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 [67]. This could be done by ranking agents' options in terms of how much aggregate welfare their consequences have [97]. Dehghani et al. [28] specify this with the principle of Utilitarianism by selecting the choice with the highest utility. Ajmeri et al. [2] choose to operationalise the principle of Prioritarianism by improving the minimum experience in the consequences of an action. Another way consequences are used is in operationalising the principle of Envy-Freeness, which Sun et al. [98] address by promoting the outcome with the lowest levels of envy between groups or individuals. Other principles, such as the Doctrine of Disparate Impact look at the representation of groups in consequences and posit that a satisfactory outcome would have equal or proportional treatment [82]. However, there are issues that arise in predicting all of the possibilities that an action could produce as this could be computationally challenging, requiring complex calculations [48]. 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 non-deterministic and probabilistic environments.

5 GAPS IN OPERATIONALISING ETHICAL PRINCIPLES

We now examine existing gaps in ethics and fairness research in computer science and AI literature, specifically in relation to implementing multiple ethical principles in reasoning capacities.

5.1 Expanding the Taxonomy

Key gaps include a lack of research on lesser-utilised principles. We suggest that future research directions include these less commonly seen principles, or incorporate a wider array of principles. Not only would this improve capacities for ethical reasoning but it could also aid the explainability of AI agents, which is important to achieve transparency so that machines are human-interpretable [16]. When looking at why an agent made a particular decision, one could refer to the exact principle they used in their explanation. This includes researching principles from other cultures outside of the Western doctrine, which is important as ethics is culturally sensitive [51]. Implementing ethical principles from different cultures will aid the accessibility and fairness of AI, as it can better apply to groups of stakeholders from different backgrounds.

5.2 Implementing Ethical Principles in STS

A gap exists in implementing ethical principles in STS, as the majority of the research we identify do not explicitly tie into STS. Tolmeijer et al. [100] study how ethical principles relate to machine ethics, but do not consider the relation of ethical principles to values and norms within the context of STS. Ajmeri et al. [2] broadly reference the principles of Egalitarianism and Utilitarianism within the context of utilising values and norms for ethical reasoning, however, this research may benefit from the consideration of other ethical principles to enable broader applicability. Future research directions could thus adapt methodologies of previous works explained in Section 4 in the context of STS.

5.3 Resolving Ethical Dilemmas

There is a gap in how to address the difficulties that may arise with the implementation of individual ethical principles. Future research directions could address this gap through the use of Pluralist and Particularist approaches.

Our findings show that there are difficulties associated with every ethical principle identified. This implies that for each principle, there are situations where it leads to an unfair outcome. Ethical dilemmas are thus scenarios in which the application of an ethical principle leads to an unfair outcome, cannot support one action over another, or conflicts with another ethical principle. When a principle cannot support one action over another, dilemmas could be resolved by examining how similar decisions were made previously [8]. If no similar decisions have been made previously, an action is selected at random. However, relying on random choice may not result in the most ethically appropriate action. Alternatively, ethical dilemmas may be mitigated through the use of Pluralist approaches, in which a variety of principles can be weighed against one another to find the fairest answer. To aid this the use of Particularism, which incorporates relevant contextual factors in ethical reasoning to identify if a certain feature is morally relevant or not [100], could help identify which principle is the most appropriate in that setting.

6 CONCLUSION

To better address the pursuit of responsible AI, research must be human-centred [34]. 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 [21]. However, dilemmas can arise when values conflict [78]. To resolve these dilemmas in satisfactory ways which promote a higher goal of fairness, ethical principles can help to determine the moral permissibility of actions [69, 73].

In this survey, we identify a broad variety of ethical principles which have been previously operationalised in computer science and AI literature (Section 3). 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 (Section 4). We find that previous literature did not operationalise multiple ethical principles in STS. Key gaps that highlight future research directions include expanding the taxonomy, implementing principles in STS, and resolving ethical dilemmas where principles conflict or lead to unfair outcomes (Section 5). We envision that the findings of this survey will contribute to the development of responsible AI by aiding the incorporation of ethical principles in reasoning capacities.

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] 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

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

[5] 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.

[6] 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.

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

[8] 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

[9] 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

[10] 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 MultiAgent Systems (AAMAS).* International Foundation for Autonomous Agents and Multiagent Systems, São Paulo, 96–104. https://doi.org/10.5555/3091125.3091145

[11] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle Friedler and Christo Wilson (Eds.). PMLR, New York, 149–159. https://proceedings.mlr.press/v81/binns18a.html

[12] 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 MultiAgent Systems (AAMAS).* International Foundation for Autonomous Agents and Multiagent Systems, London, 1–5. https://doi.org/10.5555/3463952.3463954

[13] 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

[14] Britannica. 2021. Teleological Ethics. https://www.britannica.com/topic/teleological-ethics. Accessed: 2021-09-23.

[15] 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

[16] Roberta Calegari and Federico Sabbatini. 2023. The PSyKE Technology for Trustworthy Artificial Intelligence. In *AIxIA 2022 – Advances in Artificial Intelligence*, Agostino Dovier, Angelo Montanari, and Andrea Orlandini (Eds.). Springer International Publishing, Cham, 3–16.

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

[18] Violet Chen and John Hooker. 2020. A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism. In *Proceedings of the 2020 AAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, New York, 221–227. https://doi.org/10.1145/3375627.3375844

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

[20] 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

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

[22] Gerald Allan Cohen. 1989. On the Currency of Egalitarian Justice. *Ethics* 99, 4 (1989), 906–944. https://doi.org/10.2307/2381239

[23] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, Singapore, 1106–1114.

[24] 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.

[25] David Cummiskey. 1990. Kantian Consequentialism. *Ethics* 100, 3 (1990), 586–615. https://doi.org/10.2307/2381810

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

[27] 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

[28] 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

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

[30] 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

[31] 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] https://arxiv.org/abs/2011.03108

[32] Virginia Dignum. 2017. Responsible Artificial Intelligence: Designing AI for Human Values. *ITU Journal* 1, 1 (2017), 1–8. Issue 1. http://hdl.handle.net/123456789/2181

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

[34] Virginia Dignum and Frank Dignum. 2020. Agents Are Dead. Long Live Agents!. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, Auckland, 1701–1705. https://doi.org/10.5555/3398761.3398957

[35] 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 https://arxiv.org/abs/2101.11775

[36] 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.

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

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

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

[44] 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

[45] 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

[46] 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)*. IJCAI, Melbourne, 4722–4730. https://doi.org/10.24963/ijcai.2017/658

[47] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the Engineering of Virtuous Machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, Honolulu, USA, 29—-35. https://doi.org/10.1145/3306618.3314256

[48] 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, Snowbird, 4147–4151. https://doi.org/10.5555/3016387.3016503

[49] 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

[50] 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 http://arxiv.org/abs/1907.07892

[51] 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

[52] 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

[53] 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.

[54] 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

[55] 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.

[56] Shelly Kagan. 1998. *Equality and Desert.* Oxford University Press, Oxford, Chapter 30, 298–314.

[57] 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

[58] 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

[59] 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 https://arxiv.org/abs/2109.07906

[60] Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *J. Artif. Int. Res.* 70 (may 2021), 871—-890. https://doi.org/10.1613/jair.1.12481

[61] 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

[62] Nadin Kökciyan and Pınar Yolum. 2020. TURP: Managing Trust for Regulating Privacy in Internet of Things. *IEEE Internet Computing* 24, 6 (2020), 9–16.

[63] 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

[64] Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. Association for Computing Machinery, New York, 86–92. https://doi.org/10.1145/3375627.3375808

[65] 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 (06 2021), 529–544. https://doi.org/10.1007/s43681-021-00067-y

[66] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* 55, 9, Article 177 (jan 2023), 46 pages. https://doi.org/10.1145/3555803

[67] 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

[68] Patrick Lin, Keith Abney, and George Bekey. 2011. Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence* 175, 5 (2011), 942–949. https://doi.org/10.1016/j.artint.2010.11.026 Special Review Issue.

[69] Felix Lindner, Robert Mattmüller, and Bernhard Nebel. 2019. Moral Permissibility of Action Plans. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 7635–7642. https://doi.org/10.1609/aaai.v33i01.33017635

[70] 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 MultiAgent Systems (AAMAS)*. IFAAMAS, London, 799–808. https://doi.org/10.5555/3463952.3464048

[71] Ángeles Manjarrés, Celia Fernádez-Aller, Maite López-Sánchez, Juan Antonio Rodríguez-Aguilar, and Manuel Sierra Castañer. 2021. Artificial Intelligence for a Fair, Just, and Equitable World. *IEEE Technology and Society Magazine* 40,

1 (2021), 19–24. https://doi.org/10.1109/MTS.2021.3056292
[72] Kirsten Martin. 2022. *Ethics of Care as Moral Grounding for AI.* Auerbach Publications, Boca Raton, 1–6. https://doi.org/10.1201/9781003278290
[73] 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.
[74] John S. Mill. 1863. *Utilitarianism.* Longmans, Green and Company.
[75] 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
[76] Jessica Morley, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. 2021. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines* 31 (June 2021), 239–256. https://doi.org/10.1007/s11023-021-09563-w
[77] 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.
[78] 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.
[79] 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
[80] 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
[81] Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. 2021. Ethical Online AI Systems Through Conscientious Design. *IEEE Internet Computing* 25, 6 (2021), 58–64. https://doi.org/10.1109/MIC.2021.3098324
[82] 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
[83] John Rawls. 1967. Distributive Justice. *Philosophy, Politics and Society* 1 (1967), 58–82.
[84] John Rawls. 1985. Justice as Fairness: Political not Metaphysical. *Philosophy and Public Affairs* 14, 3 (Summer 1985), 223–251. https://www.jstor.org/stable/2265349
[85] 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.
[86] Manel Rodriguez-Soto, Marc Serramia, Maite Lopez-Sanchez, and Juan Antonio Rodriguez-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
[87] Stuart J. Russell and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach.* Pearson Education Limited.
[88] 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
[89] Conrad Sanderson, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon Whittle. 2022. Towards Implementing Responsible AI. https://doi.org/10.48550/ARXIV.2205.04358
[90] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20).* Association for Computing Machinery, New York, NY, USA, 153—-158. https://doi.org/10.1145/3375627.3375804
[91] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919.
[92] Amartya Sen. 1992. *Inequality reexamined.* Clarendon Press, Oxford.
[93] 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
[94] Walter Sinnott-Armstrong. 2021. Consequentialism. In *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford.
[95] 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/
[96] 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

[97] 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.
[98] 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
[99] Justin Svegliato, Samer B. Nashed, and Shlomo Zilberstein. 2021. Ethically Compliant Sequential Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11657–11665. https://doi.org/10.1609/aaai.v35i13.17386
[100] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2021. Implementations in Machine Ethics: A Survey. *ACM Comput. Surv.* 53, 6, Article 132 (Dec. 2021), 38 pages. https://doi.org/10.1145/3419633
[101] Loïs Vanhée and Melania Borit. 2022. Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence. *J. Artif. Int. Res.* 73 (2022), 619–631.
[102] 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
[103] 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 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, Honolulu, HI, USA, 195—-200. https://doi.org/10.1145/3306618.3314289
[104] 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.
[105] 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 & SOCIETY* 1 (Nov. 2022), 1–12. https://doi.org/10.1007/s00146-022-01607-8
[106] 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.
[107] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. 2022. *AI and Ethics—Operationalizing Responsible AI*. Springer International Publishing, Cham, 15–33. https://doi.org/10.1007/978-3-030-72188-6_2

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 select 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 first 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, leading to the identification of the primary studies. From the research works gathered in this initial search, relevant citations that met the criteria 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.

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 'Multiagent' OR

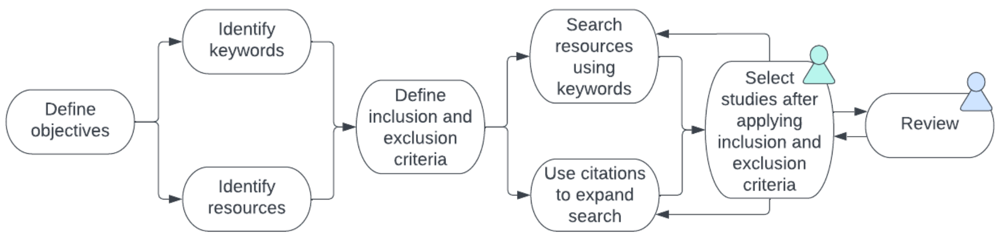

Fig. 2. Search strategy in brief.

'Multiple-User') AND ('Responsible' OR 'Fairness' OR 'Consequentialism' OR 'Deontology' OR 'Egalitarianism' OR 'Equality' OR 'Ethics' OR 'Utilitarianism').

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 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 works related to multiple-user social dilemmas to examine how ethical principles are operationalised in these settings. We exclude studies about how ethical principles affect other non-social dilemmas. Fifth, 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). Sixth, we include studies about normative ethical principles and AI, but we exclude studies solely about AI principles. This is because, whilst AI principles contain important information about ethical implementation, it is out of the scope of this review. Seventh, 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.

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 the types of research that they were (seen in Table 1). 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 computer science (e.g., [64]). 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., [98]). Some works had a mixture of both identification and implementation (e.g., [60]). 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.

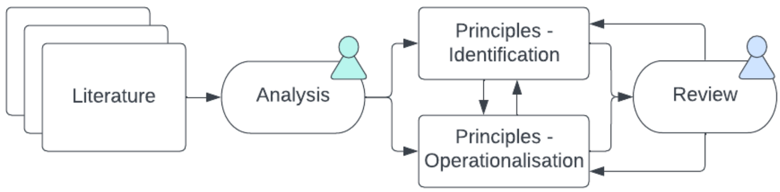

Fig. 3. Methodology to extract principle identification and operationalisation from literature.

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 the external and internal validity of the study. The 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. The initial search string is based on preliminary research, and as the review continues more key terms are identified. To address this concern, it is ensured that the aims of the review are carefully scoped which allows for the identification of a good array of initial relevant terms. 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 Kitchenhams's [61] 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.

Table 6. Technical Implementation of Principles.

Implementation Type		Ethical Principles						
		Consequentialism	Utilitarianism	Maximin	Envy-Freeness	Doctrine of Double Effect	Doctrine of Disparate Impact	Do No Harm
Logical Reasoning	Deductive Logic	–	[85]	–	–	–	–	
	Non-Monotonic Logic	–	[10, 67]	[10]	–	[10]	–	–
	Abductive Logic	–	–	–	–	[75]	–	–
	Deontic Logic	–	[69]	–	–	[46, 69]	–	[69]
	Rule-Based Systems	[23]	[8, 28, 85]	[2]	–	–	–	[30]
	Event Calculus	–	[10, 67]	[10]	–	[10, 47]	–	–
	Knowledge Representation and Ontologies	[23]	[28, 85]	–	–	–	–	–
	Inductive Logic	–	–	–	–	–	–	–
Probabilistic Reasoning	Bayesian Approaches	–	[7]	–	–	–	–	–
	Markov Models	[86]	[99]	–	–	[46]	–	–
	Statistical Inference	–	–	–	–	–	–	–
Learning	Decision Tree	–	[8]	–	–	–	–	–
	Reinforcement Learning	[86]	–	[2]	–	–	–	–
	Inverse Reinforcement Learning	–	–	–	–	–	–	–
	Neural Networks	–	[52]	–	–	–	–	–
	Evolutionary Computing	–	–	–	–	–	–	–
Optimisation		–	[6–8, 19, 64]	[19, 31, 64, 82]	[98]	–	[11, 82]	–
Case-Based Reasoning		–	[28]	–	–	–	–	–

Table 8. 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
Multiple-user social dilemmas
Normative ethics and multiple-user AI
Normative ethics and STS
Normative ethical principles and AI
Bias when related to ethical principles

Exclusion
Works about meta-ethics or applied ethics outside of computer science
Studies about specific ML methodology
Non-social dilemmas
Studies of multiple-user AI that do not mention ethics
Studies of STS that do not mention ethics
AI principles
Studies about bias without reference to ethical principles