



Woodgate, J., & Ajmeri, N. (2022). *Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions*.  
<https://doi.org/10.48550/arXiv.2208.12616>

Early version, also known as pre-print

Link to published version (if available):  
[10.48550/arXiv.2208.12616](https://doi.org/10.48550/arXiv.2208.12616)

[Link to publication record in Explore Bristol Research](#)  
PDF-document

This is a pre-print server version of the article. It first appeared online via arXiv at <https://doi.org/10.48550/arXiv.2208.12616>. Please refer to any applicable terms of use of the publisher.

## University of Bristol - Explore Bristol Research

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Woodgate & Ajmeri

# Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions

Jessica Woodgate

YP19484@BRISTOL.AC.UK

Nirav Ajmeri

NIRAV.AJMERI@BRISTOL.AC.UK

*Department of Computer Science, University of Bristol, Bristol, United Kingdom, BS8 1UB*

## Abstract

The rapid adoption of artificial intelligence (AI) necessitates careful analysis of its ethical implications. In addressing ethics and fairness implications, it is important to examine the whole range of ethically relevant features rather than looking at individual agents alone. This can be accomplished by shifting perspective to the systems in which agents are embedded, which is encapsulated in the macro ethics of sociotechnical systems (STS). Through the lens of macro ethics, the governance of systems — which is where participants try to promote outcomes and norms which reflect their values — is key. However, multiple-user social dilemmas arise in an STS when stakeholders of the STS have different value preferences or when norms in the STS conflict. To develop equitable governance which meets the needs of different stakeholders, and resolve these dilemmas in satisfactory ways with a higher goal of fairness, we need to integrate a variety of normative ethical principles in reasoning. Normative ethical principles are understood as operationalizable rules inferred from philosophical theories. A taxonomy of ethical principles is thus beneficial to enable practitioners to utilise them in reasoning.

This work develops a taxonomy of normative ethical principles which can be operationalized in the governance of STS. We identify an array of ethical principles, with 25 nodes on the taxonomy tree. We describe the ways in which each principle has previously been operationalized, and suggest how the operationalization of principles may be applied to the macro ethics of STS. We further explain potential difficulties that may arise with each principle. We envision this taxonomy will facilitate the development of methodologies to incorporate ethical principles in reasoning capacities for governing equitable STS.

## 1. Introduction

The rapid development of AI systems entails the importance of understanding their ethical impact (Dastani & Yazdanpanah, 2022). The recent shift in agent research from emphasis on single agents to multiagent systems (MAS: multiple technical agents deployed into a common environment, Rădulescu et al., 2019) necessitates careful analysis of the ethical implications of MAS (Dignum & Dignum, 2020; Chopra & Singh, 2018). To develop these systems with ethics in mind, fairness is key (Floridi & Cowls, 2019). Fairness is understood as non-discrimination, where discrimination is prejudice against people based on sensitive attributes (Bishr, 2018; Mehrabi et al., 2021). In pursuing the development of a fair MAS, shifting perspective to sociotechnical systems (STS) is important as it incorporates the human element in ethical reasoning (Murukannaiah & Singh, 2020). In an STS, humans and agents work together as ethical duos, with the agent acting on behalf of their human counterpart. Within the context of an STS, it is also important to adopt the perspective of

arXiv:2208.12616v1 [cs.CY] 12 Aug 2022

macro ethics (Chopra & Singh, 2018). Macro ethics focusses on the governance of the STS and addresses the full range of ethically relevant features, as opposed to micro ethics, which focusses on the more restricted perspective of a single agent’s decision-making within an STS. The scope of our work therefore lies in the macro ethics perspective of STS, considering the governance of systems.

The governance of STS is when stakeholders try to promote outcomes and norms (rules of expected behaviour, Morris-Martin et al., 2019) that align with their values (what is important to us in life, Schwartz, 2012). This is important because ethics should be understood as a reflective development process that incorporates context (Kökciyan & Yolum, 2020; Morley et al., 2021; Manjarrés et al., 2021; Zhu et al., 2022). Values and norms are thus pivotal for ethical reasoning (Ajmeri et al., 2020; Dignum et al., 2018; Singh, 2013; Yazdanpanah et al., 2021). However, users may have different value preferences or their values may conflict with norms (Dechesne et al., 2013). Challenges thus arise in making decisions concerning multiple users (Kökciyan et al., 2017; Liao et al., 2019). These scenarios are known as multiple-user social dilemmas and can occur in mundane settings, for instance, a smart home agent deciding when to put the heating on, taking into account the preferences of existing users and other contextual features.

Resolving these dilemmas in satisfactory ways with an overarching goal of fairness may be aided by the incorporation of normative ethical principles in reasoning (Woodgate & Ajmeri, 2022). Normative ethics is the study of practical means to determine the ethicality of an action through the use of principles and guidelines, or the rational and systematic study of the standards of right and wrong (Murukannaiah & Singh, 2020). By examining how normative ethical theories have been used to improve fairness considerations, and how they have been operationalized to make ethical decisions in artificial intelligence (AI), it may be possible to implement normative ethical theories to make ethical decisions in MAS that have the overall goal of fairness. Normative ethical principles have previously been utilised in the context of choosing fairness metrics for binary ML algorithms in works such as Binns (2018) and Leben (2020). The implementation of normative ethical principles in the decision making of agents (acting entities that perform actions to achieve goals, which are decisions made using AI Pdamkar, 2021) has been used to enable agents to make ethical judgements in specific contexts (Cointe et al., 2016). They can also be applied to improve fairness considerations in systematic analysis (Saltz et al., 2019; Conitzer et al., 2017).

## 1.1 Motivation for a Taxonomy of Ethical Principles

The motivation for this work therefore stems from the need to improve fairness considerations in MAS. Fairness can be improved through the operationalization of normative ethical principles in the governance of STS (Woodgate & Ajmeri, 2022). Ethical principles imply certain logical propositions that must be true in order for a given action plan to be ethical (Kim et al., 2021). Therefore, the application of ethical principles may be useful in order to methodically think through dilemmas and promote satisfactory outcomes (Conitzer et al., 2017). Such principles can help to guide normative judgements, understand different perspectives, and determine the moral permissibility of concrete courses of actions (Canca, 2020; McLaren, 2003; Saltz et al., 2019; Lindner et al., 2019).

Ethical thinking needs to be fostered through the appreciation of a variety of different approaches, considering the strengths and limitations of each (Burton et al., 2017; Robbins & Wallace, 2007). We envision that a taxonomy of ethical principles will aid this ethical thinking.

## 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 (2020), and (2) those adapted from other disciplines like medicine and bioethics such as those suggested by Floridi and Cows (2019), Jobin et al. (2019), Fjeld et al. (2020), and Cheng et al. (2021) 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.

These two types of principles are related but distinct domains which can be easily confused. To ensure clarity of terminology, we refer to principles from normative ethics as *ethical principles*, and those as highlighted by Floridi and Cows (2019) and Jobin et al. (2019) as *AI keystones*.

**AI Keystones** 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** Operationalizable rules inferred from philosophical theories such as Deontology and Consequentialism.

Existing taxonomies and surveys are present in the relevant but distinct domain of AI keystones such as Jobin et al. (2019), Floridi and Cows (2019) and Khan et al. (2021), however, not in ethical principles as is defined here. The work of Tolmeijer et al. (2021) has much relevant information, however, the authors do not capture the whole range of ethical principles we capture. In addition, Tolmeijer et al. look at ethics from the perspective of machine ethics, rather than AI ethics and how it relates to MAS, as we aim to address. Similarly, Yu et al. (2018) identify a high-level overview of ethical principles, but fail to recognize the extent that has been found in our work, and do not consider them in the same level of depth. Dignum (2019), Leben (2020), and Robbins and Wallace (2007) give summaries of normative ethics, however, they have not considered principles such as Do No Harm, referred to by Linder et al. (2019). To enable broader applicability, these works may benefit from a formal taxonomy including other ethical principles seen in computer science.

## 1.3 Objective and Contributions

Our broad objective is to investigate the current understanding of ethical principles in AI and computer science and how these principles are operationalized. Specifically, we address the following questions:

**Q<sub>p</sub> (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.

**Q<sub>o</sub> (Operationalization).** How have ethical principles been operationalized in AI and computer science research?

This question looks at the identified principles to examine how they have been operationalized in AI and computer science. Works such as Leben (2020) and Tolmeijer et al. (2021) give some guidance as to how certain ethical principles may be operationalized, however, they are not extensive and miss some principles.

**Q<sub>g</sub> (Gaps).** What are existing gaps in ethics and fairness research in AI and computer science, specifically in relation to operationalized principles in an STS?

This question aids analysis of the gaps that exist in operationalizing the principles within the scope of STSs to direct future research.

## 1.4 Organisation

Section 2 explains the methodology in brief, threats to validity that arose. 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, evaluating how each objective has been answered and highlighting existing gaps. Section 4 includes details of each Deontological ethical principle, including how they have been previously operationalized and potential difficulties that may arise. Section 5 does the same for Teleological principles. Section 6 concludes with our key takeaways and directions with future work.

## 2. Methodology in Brief for Reproducibility

Taking inspiration from software engineering research for reproducibility and extendability of the taxonomy, we follow Kitchenham’s (2007) guidelines on conducting a systematic literature review to develop our taxonomy for ethical principles. Figure 1 outlines our method in brief.

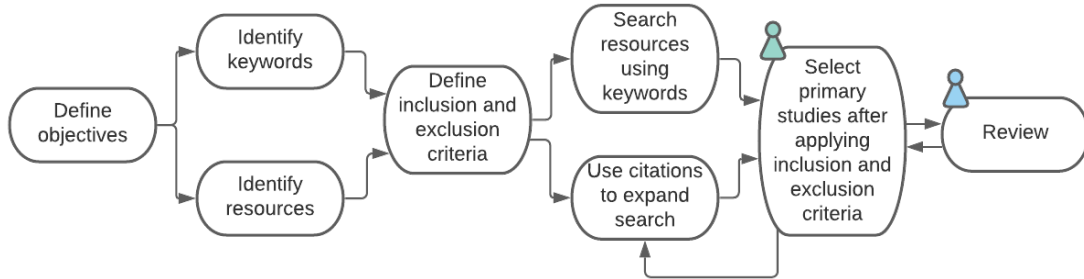


Figure 1: Method in brief

After defining our objective and questions, we formed the strategy to search for primary studies by identifying keywords and resources. We select Google Scholar and the University of Bristol Online Library as resources to search. 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. The search string used was ('AI' OR 'Agent' OR 'ML' OR 'Multiple-User' OR 'Multiagent') AND ('Bias' OR

‘Consequentialism’ OR ‘Deontology’ OR ‘Egalitarianism’ OR ‘Equality’ OR ‘Ethics’ OR ‘Utilitarianism’).

After inspecting up to the first 5 pages of results in each resource, we narrowed the search by applying the inclusion and exclusion criteria to the titles, eliminating obviously irrelevant studies. 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.

## 2.1 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 excluding works outside of these resources, and also by following relevant citations from primary studies to expand the scope. 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 related to individual or group fairness. We exclude works about fairness in specific ML methodology, as that is outside the scope of this project. Third, we include works related to multiple-user social dilemmas in order to examine how ethical principles are operationalized in these settings. We exclude studies about how ethical principles affect other non-social dilemmas. Fourth, we include the intersection of normative ethics and multiple-user AI or MAS research, whereas we exclude non-ethical studies (e.g. about technical implementation) in this area. Fifth, we include studies about normative ethical principles and AI, but we exclude studies solely about AI keystones. This is because, whilst AI keystones contain important information about ethical implementation, it is out of the scope of this review. Sixth, we include studies about bias when related to ethical principles, as this is relevant to how ethical principles affect fairness, however we exclude studies about bias that do not talk about ethical principles.

## 2.2 Relevant Works

We conducted an initial search on 01-Jun-21. 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 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 54 papers being included in the review. We conducted a second search on 23-May-22, resulting in a further 6 papers being included in the review. Based on our review, we create a taxonomy of ethical principles via an iterative process, adding nodes when new principles were identified in literature and revising the structure accordingly.

### 3. Ethical Principles in Computer Science and AI

Research on AI and ethical principles is broadly categorised into twelve key principles, based on their definition of normative ethics, and five types of study, based on the structure and contributions of the paper. Tables 1 and 2 map out our findings.

Within ethics, there are two main strands of theory: Deontology and Teleology. Deontological theories revolve around rules, rights, and duties (Murukannaiah & Singh, 2020; Wallach et al., 2008). Teleological ethics, on the other hand, derive duty or moral obligation from what is good or desirable as an end to be achieved, (Britannica, 2021). Teleological ethics can be further divided into Consequentialism, Egoism, and Virtue Ethics (SOAS, 2021). Figure 2 displays the taxonomy of principles identified in literature in a tree structure, mapping out how they relate to each other.

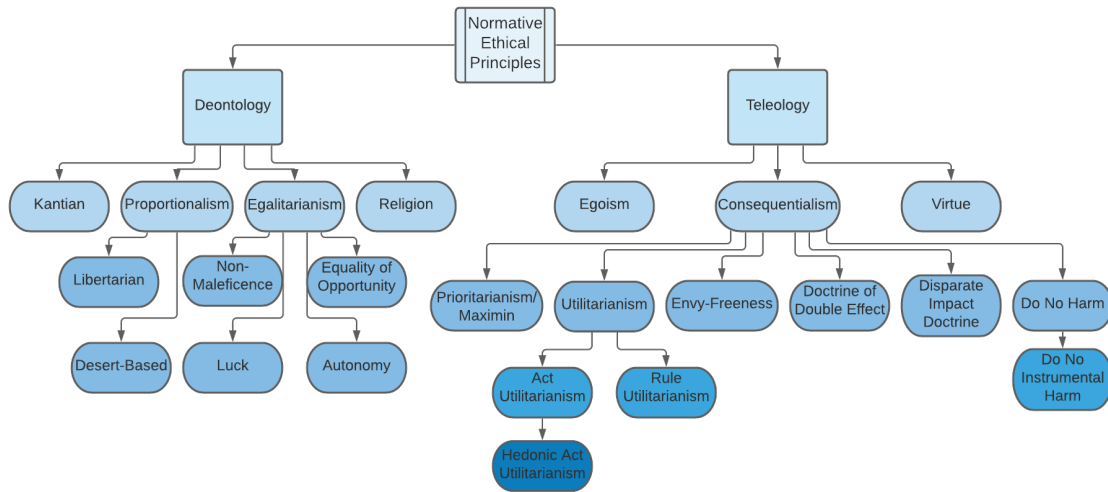


Figure 2: Taxonomy of Ethical Principles

We find that certain principles, such as Utilitarianism, were much more discussed than other principles such as Do No Harm. This is perhaps because Utilitarianism is a historically prolific theory, therefore it is well-known and more likely to be utilised by researchers.

We also find that there is a significant 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. (2016), Greene et al. (2016), and Anderson and Anderson (2014). These works would perhaps benefit by more clearly specifying the ethical principles they are using, in order to allow for more precise operationalization.

#### 3.1 Operationalization of Ethical Principles

We iterated over the papers identified in our review to conduct analysis of previous operationalization of ethical principles. We find that the architecture must be defined as to whether the principles are integrated into reasoning capacities in a top-down, bottom-up, or hybrid approach. In addition to this, a definition of welfare is necessary in order to

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

Topic \ Type	Frameworks (Conceptualization)	Frameworks (Application)
Deontology	(Abney, 2011; Binns, 2018; Brink, 2007; Cointe et al., 2016; Greene, Rossi, Tasioulas, Venable, & Williams, 2016; Leben, 2020; Murukannaiah & Singh, 2020; Saltz et al., 2019; Wallach, Allen, & Smit, 2008)	(Anderson & Anderson, 2014; Berreby, Bourgne, & Ganascia, 2017; Dehghani, Tomai, & Klenk, 2008; Honarvar & Ghasem-Aghaee, 2009; Limarga, Pagnucco, Song, & Nayak, 2020; Lindner et al., 2019; Robbins & Wallace, 2007)
Egalitarianism	(Binns, 2018; Cohen, 1989; Dworkin, 1981; Fleurbaey, 2008; Friedler, Scheidegger, & Venkatasubramanian, 2021; Leben, 2020; Murukannaiah, Ajmeri, Jonker, & Singh, 2020; Rawls, 1985; Sen, 1992)	(Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012)
Proportionalism	(Etzioni & Etzioni, 2016; Kagan, 1998; Leben, 2020)	(Dwork et al., 2012)
Kantian	(Abney, 2011; Kant, 2011; Hagerty & Rubinov, 2019; Kim et al., 2021; Wallach et al., 2008)	(Berreby et al., 2017; Limarga et al., 2020; Robbins & Wallace, 2007)
Virtue	(Abney, 2011; Anderson & Anderson, 2007; Brink, 2007; Cointe et al., 2016; Greene et al., 2016; Hagerty & Rubinov, 2019; Murukannaiah & Singh, 2020; Saltz et al., 2019; Wallach et al., 2008)	(Govindarajulu, Bringsjord, Ghosh, & Sarathy, 2019; Honarvar & Ghasem-Aghaee, 2009; Robbins & Wallace, 2007)
Consequentialism	(Abney, 2011; Brink, 2007; Cointe et al., 2016; Cummiskey, 1990; Greene et al., 2016; Hagerty & Rubinov, 2019; Leben, 2020; Saltz et al., 2019; Sinnott-Armstrong, 2021; Suikkanen, 2017)	(Berreby et al., 2017; Limarga et al., 2020)
Utilitarianism	(Abney, 2011; Anderson & Anderson, 2007; Brink, 2007; Honarvar & Ghasem-Aghaee, 2009; Kim et al., 2021; Leben, 2020; Mill, 1863; Murukannaiah et al., 2020; Murukannaiah & Singh, 2020; Wallach et al., 2008)	(Ajmeri et al., 2020; Anderson, Anderson, & Armen, 2004; Berreby et al., 2017; Dehghani et al., 2008; Limarga et al., 2020; Lindner et al., 2019; Robbins & Wallace, 2007)
Maximin	(Leben, 2020; Rawls, 1967)	(Ajmeri et al., 2020)
Envy-Freeness	(Boehmer & Niedermeier, 2021)	–
Doctrine of Double Effect	–	(Berreby et al., 2017; Govindarajulu & Bringsjord, 2017; Lindner et al., 2019)
Doctrine of Disparate Impact	(Binns, 2018)	–
Do No Harm	(Dennis, Fisher, Slavkovik, & Webster, 2016)	(Lindner et al., 2019)



Table 2: Categorisation of Research Reviewed with Principles Extracted: Algorithm and Viewpoint or Review Studies

Topic \ Type	Algorithms	Viewpoint or Review
Deontology	(Rodriguez-Soto, Serramia, Lopez-Sanchez, & Rodriguez-Aguilar, 2022)	(Kazim & Koshiyama, 2020; Hagedorff, 2020; Tolmeijer et al., 2021; Yu et al., 2018)
Egalitarianism	–	(Lee, Floridi, & Singh, 2021)
Proportionalism	–	–
Kantian	–	(Tolmeijer et al., 2021; Kumar & Choudhury, 2022)
Virtue	(Rodriguez-Soto et al., 2022)	(Kazim & Koshiyama, 2020; Hagedorff, 2020; Tolmeijer et al., 2021; Vanh�e & Borit, 2022; Yu et al., 2018)
Consequentialism	(Rodriguez-Soto et al., 2022)	(Etzioni & Etzioni, 2017; Tolmeijer et al., 2021; Yu et al., 2018)
Utilitarianism	(Rodriguez-Soto et al., 2022)	(Etzioni & Etzioni, 2017; Kazim & Koshiyama, 2020; Yu et al., 2018; Kumar & Choudhury, 2022)
Maximin	(Diana, Gill, Kearns, Kenthapadi, & Roth, 2021; Sun, Chen, & Doan, 2021)	(Lee et al., 2021)
Envy-Freeness	(Sun et al., 2021)	(Lee et al., 2021)
Doctrine of Double Effect	–	(Deng, 2015)
Doctrine of Disparate Impact	(Patel, Khan, & Louis, 2020)	–
Do No Harm	–	–

understand what the ‘good’ is understood as, or what the principle is aiming for in its application. The operationalization of principles in reasoning largely divides into three camps in which actions are chosen either according to: (1) how the action adheres to certain rules, (2) by evaluating the consequences the action produces, or (3) through the development of virtues. There are also various factors that depend on particular principles as to whether they are necessary for that principle to be implemented.

### 3.1.1 CLARIFYING THE ARCHITECTURE

**Bottom-Up** *Machines learning to make ethical decisions through observation, without being taught any formal rules (Etzioni & Etzioni, 2017).*

**Top-Down** *Ethics is installed directly into machines (Kim et al., 2021) as rules that dictate what the morally correct action is (Lin et al., 2011).*

**Hybrid** *Incorporate both ethical reasoning and empirical observation, supplementing top-down imposition of rules with bottom-up observation of context (Berreby et al., 2017).*

In order to engineer morally sensitive systems, practitioners must decide the architecture for integrating ethical theory (Wallach et al., 2008). These fall within two broad approaches: the top-down imposition of ethical theories, and the bottom-up building of systems with goals that may or may not be explicitly specified.

Bottom-up approaches are understood as machines learning to make ethical decisions by observing human behaviour in actual situations, without being taught any formal rules or moral philosophy (Etzioni & Etzioni, 2017). While bottom-up approaches typically ask the machine to learn prescriptive norms from experience, top-down approaches install ethics directly into the machine (Kim et al., 2021). Top-down approaches are rule-based: ethics is understood as the investigation of right actions through identifying rules that should be followed in order to perform the morally correct (or at least permissible) action (Lin et al., 2011). In addition to these there are hybrid approaches, which incorporate both ethical reasoning and empirical observation.

We find that many works (Limarga et al., 2020; Sun et al., 2021; Diana et al., 2021; Dehghani et al., 2008) used top-down approaches to integrate ethical principles into the reasoning capacities of machines. Other works such as Berreby et al. (2017) implemented hybrid approaches, where top-down imposition of rules is supplemented with bottom-up observation of contextual information. We find no works that use ethical principles in a purely bottom-up approach. Future research could extrapolate ethical principles from experience. However, a difficulty with extrapolation is how to formalize what ethical principles would be if agents do not have prior knowledge or definitions of them.

### 3.1.2 DEFINING WELFARE

**Welfare** *The assessment of what is good or valuable, and what constitutes a satisfactory outcome (Binns, 2018).*

For all principles, Deontological or Teleological, we find that having a definition of welfare was necessary in order to implement them. This is because welfare must be considered in the pursuit of fairness (Fleurbaey, 2008). Welfare could be pleasure or preference satisfac-

tion (Cohen, 1989), income and assets (Rawls, 1958; Dworkin, 1981), or the ability and resources necessary to do certain things (Sen, 1992). It is the assessment of what is good or valuable, and what constitutes a satisfactory outcome (Binns, 2018). In other words, it is what is being aimed for in the application of a certain principle. The definition of welfare could seem abstract, and vary between contexts and cultures. Consistent and fair methods are thus needed for determining welfare. An example of how welfare has been used to operationalize Egalitarianism (which states that humans are in some fundamental sense equal and efforts should be made to correct forms of inequality, Binns, 2018), is in Murukannaiah et al. (2020). Here, the authors suggest that the principle entails maximizing disparity across stakeholders with respect to satisfying their preferences. Welfare in this instance is thus understood as preference satisfaction, and when applying Egalitarian principles, equal preference satisfaction is what is being aimed for.

### 3.1.3 USING RULES, CONSEQUENCES, OR VIRTUES

We find that previous works have operationalized principles either through the application of rules and then choosing an action based on how it accords to certain rules, by evaluating consequences and then choosing an action based on the consequences it produces, or through the development of virtuous characteristics.

- *Applying Rules* Some approaches suggest operationalizing principles by applying a set of rules to possible actions to determine which ones would be satisfactory, such as Deontological implementations in Abney (2011), Greene et al. (2016), and Berreby et al. (2017). Examples of this would be applying the rule that the disparity of preference satisfaction for stakeholders should be maximised, extracted from the principle of Egalitarianism (Murukannaiah et al., 2020). Another example is applying the rule that stakeholders should be treated proportionally based on their contributions to production (Leben, 2020). However, due to the abstract nature of ethics, there are difficulties that arise in finding appropriate ways to encode ethical principles in concrete rules (Tolmeijer et al., 2021). Creating systematic ways of encoding ethical principles into rules to use in the context of STS could thus be a direction for future research.
- *Analyzing Consequences* Consequentialist principles may be operationalized by evaluating the consequences of different actions (Limarga et al., 2020). This could be done by ranking agents' options in terms of how much aggregate welfare their consequences have (Suikkanen, 2017). Dehghani et al. (2008) specify this with the principle of Utilitarianism by selecting the choice with the highest utility. Instead of choosing the consequence with the most welfare, Ajmeri et al. (2020) choose to operationalize the principle of Prioritarianism by improving the minimum experience in the consequences of an action. Another way consequences are used is in operationalizing the principle of Envy-Freeness, in which Sun et al. (2021) promote the outcome with the lowest levels of envy between groups or individuals. Other principles, such as the Disparate Impact Doctrine look at the representation of groups in consequences and posit that a satisfactory outcome would have equal or proportional treatment (Patel et al., 2020). However, there are issues that arise in predicting all of the possibilities that an action could produce (Greene et al.,

2016). The simulation of consequences in the context of reasoning capacities used to govern STS could therefore be a direction for future research.

- *Developing Virtues* For Virtue Ethics, ethicality stems from the inherent character of an individual (Murukannaiah & Singh, 2020; Wallach et al., 2008; Brink, 2007; Kazim & Koshiyama, 2020; Yu et al., 2018; Abney, 2011). To solve a problem according to this theory, virtuous characteristics should be applied (Robbins & Wallace, 2007). Thus, the theory can be operationalized through the instantiation of virtues (Tolmeijer et al., 2021). This is exemplified in Govindarajulu et al. (2017) where virtues are instantiated by using computational formal logic to formalize emotions, represent traits, and establish a process of learning traits. Therefore, the development of virtuous characteristics is needed in order to implement Virtue Ethics into machines. This has been done in previous literature by utilising computational formal logic to establish virtuous traits, and those machines that have virtuous characteristics will act ethically by definition. However, there are difficulties in that Virtue Ethics can be difficult to apply to individual situations (Saltz et al., 2019), and there are thus challenges that arise with the application of Virtues across time and culture (Tolmeijer et al., 2021). Future research could therefore incorporate the applicability of Virtue Ethics across different contexts in MAS.

#### 3.1.4 CONSIDERING OTHER INPUTS

Lastly, we identify various factors that vary from principle to principle as to whether they were necessary for operationalization. These inputs may be difficult to obtain as they can seem subjective, especially the inputs of contribution and autonomy. Further research into analysing the role of these inputs would therefore be helpful, and also to understand if there are any additional inputs not identified here.

- *Luck* In order to apply the principle of Luck Egalitarianism, levels of luck must be considered as it entails no-one should end up worse off due to bad luck (Lee et al., 2021). This means that people should not be made worse or better off because of factors that are outside of their control, for example circumstances of birth.
- *Autonomy* Autonomy, understood as the extent to which an agent's actions are free (actions for which the agent has reasons, Kim et al., 2021), is a necessary input for the principle of Egalitarian Autonomy. This principle denotes that a minimum level of autonomy should be attained and equally distributed (Fleurbaey, 2008). For a satisfactory outcome, it is thus necessary to understand existing levels of autonomy in order to apply the principle.
- *Contribution* The consideration of contribution was necessary for the implementation of certain principles such as Libertarian Proportionalism. For this principle, a satisfactory outcome may be found by evaluating each person's total contribution at the time of consent (Leben, 2020). In order to do this, it is therefore necessary to define what resource people are contributing (e.g. effort, money, time, etc.).
- *Utility* For Utilitarian principles, it is necessary to obtain a definition of utility and quantities of it in order to apply them. This is because these principles promote actions

that maximize utility, or ‘the good’ (Limarga et al., 2020). Therefore, solutions that have the highest utility will be those that align best with Utilitarian principles.

- *Envy* In applying the principle of Envy-Freeness, envy should be eliminated or minimized (Sun et al., 2021). In order to do this, levels of envy must therefore be considered in the process of decision making.

### 3.2 Gaps in Ethics and Fairness Research in Computer Science and Artificial Intelligence

We now examine existing gaps in ethics and fairness research in computer science and artificial intelligence literature, specifically in relation to implementing these principles in multiagent systems.

*Expanding the Taxonomy.* Key gaps include a lack of research on lesser-utilized principles such as the Doctrine of Double Effect and Proportionalism. We suggest that future research could include these less commonly seen principles, or incorporate a wider array of principles. Not only would this allow for agents that have better ethical reasoning capacities, but it would also aid the explainability of AI agents. When looking at why an agent made a particular decision, one could refer to the exact principle they used in their explanation. There is also an importance of researching principles from other cultures outside of the Western doctrine, and these should be incorporated into ethical reasoning and the design of ethical AI agents. This will aid the accessibility and fairness of technology, as it can better apply to groups of stakeholders from different backgrounds.

*Considering Ethical Principles in STS.* The majority of research identified did not explicitly tie in to STS. Tolmeijer et al. (2021) 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. (2020) 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 work could thus adapt methodologies suggested by these authors to the context of STS.

*Resolving Ethical Dilemmas.* Lastly, the findings show that there are difficulties associated with every ethical principle identified. This implies that for each principle, there will be some situations in which 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. This may be mitigated through the use of Pluralist approaches, in which a variety of principles can be weighed against one another in order to find the fairest answer. To aid this, the use of Particularism (the incorporation of relevant contextual factors in ethical reasoning to identify if a certain feature is morally relevant or not, Tolmeijer et al., 2021), could help to identify which principle is the most appropriate in that setting. There is thus a gap in how to address the difficulties that may arise with the implementation of a particular ethical principle, and this may be addressed in future work through the use of Pluralist and Particularist approaches.

## 4. Deontological Ethical Principles

This section examines each Deontological principle identified in the review. For this section and Section 5, each subsection is structured in a way that first explains how previous work has defined that principle, then how previous work has operationalized the principle. If no previous work operationalizing the principle was identified, we suggest how it may be operationalized in the context of an STS. We then explain potential difficulties with the application of that principle.

### 4.1 Deontology

**Deontology** *Conforming to rules, laws, and norms (Murukannaiah & Singh, 2020; Hagendorff, 2020), and respecting relevant obligations and permissions (Cointe et al., 2016) that stem from duties and rights (Wallach et al., 2008; Saltz et al., 2019; Yu et al., 2018; Kazim & Koshiyama, 2020; Robbins & Wallace, 2007).*

Deontology entails conforming to rules, laws, and norms (Murukannaiah & Singh, 2020; Hagendorff, 2020), and respecting relevant obligations and permissions (Cointe et al., 2016) that stem from duties and rights (Wallach et al., 2008; Saltz et al., 2019; Yu et al., 2018; Kazim & Koshiyama, 2020; Robbins & Wallace, 2007; Rodriguez-Soto et al., 2022). The permissibility of action, for Deontological theories, lies within the intrinsic character of the act itself; according to Deontological approaches, an action is permissible if and only if the act itself is intrinsically morally good or different, independent of the value it produces (Lindner et al., 2019; Limarga et al., 2020; Brink, 2007).

*Operationalizing Deontology.* Operationalizing Deontology may be achieved by taking a set of rules that are either implemented in design (using a top-down architecture) or acquired through learning (using a bottom-up architecture), and ensuring technology acts in accordance with them (Abney, 2011; Greene et al., 2016). These rules could be supplemented with contextual information (including a definition of welfare) and preference aggregation to obtain satisfactory outcomes. Frameworks implementing this methodology, such as Berreby et al. (2017), work by collecting contextual information to simulate the outcome of actions, and then assessing the ethical considerations of that outcome using Deontological specifications. Other works such as Limarga et al. (2020) use Deontology for automated moral reasoning. Similarly, Tolmeijer et al. (2021) portray the implementation of Deontology by inputting the action (in terms of mental states and consequences), using rules and duties as the decision criteria, and the extent to which they fit with the rule as the mechanism.

Other uses for Deontology include choosing between incompatible fairness metrics (Binns, 2018), or to evaluate distribution of binary classification algorithms (Leben, 2020). Some also suggest using Deontology only in specific circumstances: Dehghani et al. (2008) choose to implement Deontology in situations with ‘sacred values’ and without an order of magnitude difference between outcomes, using it to select the choice of action that doesn’t violate the sacred value.

*Difficulties.* There are many difficulties established with Deontology, some of which have been discussed in computer science literature. 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 and thus basic logic would be unable to adequately capture

complex ethical insights (Abney, 2011; Saltz et al., 2019). 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.

There are also issues related to which rules should be implemented. Rules are expected to be strictly followed, implying that for every exception they must be amended, which may make them rather long (Tolmeijer et al., 2021). Determining the right level of detail, is important to ensure interpretability for the machine. Another issue is when conflicts between rules occur. This may be amendable by ordering or weighing the rules, but the order of importance must still be determined, and it also assumes that all relevant rules are determined before they are used.

## 4.2 Egalitarianism

**Egalitarianism** *Humans are in some fundamental sense equal, and efforts should thus be made to mitigate inequalities (Binns, 2018).*

Egalitarianism stems from the notion that human beings are in some fundamental sense equal, and therefore efforts should be made to avoid and correct certain forms of inequality (Binns, 2018). This sometimes means that certain valuable things should be equally distributed. Egalitarianism welfare may be pleasure or preference satisfaction (Cohen, 1989), income and assets (Rawls, 1985; Dworkin, 1981), or the ability and resources necessary to do certain things (Sen, 1992). The importance of each currency of Egalitarianism may plausibly differ between contexts, and one should consider that different people may value the same outcome or set of harms and benefits differently (Binns, 2018).

*Operationalizing Egalitarianism.* As Egalitarianism is a Deontological principle, it can be operationalized through the application of rules in a top-down, bottom-up, or hybrid architecture. In applying these rules, a definition of welfare is needed. There are various ways that Egalitarianism has been implemented in literature, for example Murukannaiah et al. (2020) suggest maximizing disparity across stakeholders with respect to satisfying their preferences. In this example welfare would thus be preference satisfaction. Dwork et al. (2012) see Egalitarianism in terms of individual fairness as the principle that any two individuals who are similar with respect to a particular class should be classified similarly.

Another approach focuses on distribution of rights instead of classification of individuals. Leben (2020) argues that it means equal rights (and thus equal shares) should be conferred to each member of the population, yet if it is impossible to achieve equality across all metrics for the entire population, they suggest a distribution that minimizes the distance to some fairness standard (e.g. size of population).

*Difficulties.* Egalitarianism may be not concerned with an unequal state of affairs per se, but rather with the way in which that state of affairs was produced (Binns, 2018). A prominent debate exists as to whether a single Egalitarian calculus should be applied across different social contexts, or whether there are internal ‘spheres of justice’ in which different incommensurable logics of fairness might apply, and between which redistributions might not be appropriate.

#### 4.2.1 EGALITARIANISM: NON-MALEFICENCE PRINCIPLE

**Egalitarian Non-Maleficence** *Egalitarianism imposed across harms but not benefits; the fairest outcome is one in which harms are equally distributed (Leben, 2020).*

The non-maleficence principle imposes Egalitarianism across harms but not benefits (Leben, 2020). It is otherwise known as the compensation view. It thus emphasizes that the fairest outcome is one in which any harms are equally distributed amongst everyone (individuals or groups). Benefits may be unequally distributed, with some having more benefits than others, and this would still be fair.

*Operationalizing Non-Maleficence.* No examples of the Non-Maleficence principle being operationalized were found in included literature. As it is a Deontological principle, it could be operationalized by applying the principle as a rule: that Egalitarianism should be imposed across harms but not benefits. This could be done by a bottom-up, top-down, or hybrid approach. In order to implement it, a definition of welfare would be needed to understand what would constitute a benefit and what would be a harm. For example, if welfare is preference satisfaction then a benefit would be the satisfaction of preferences, and a harm would be the dissatisfaction of preferences.

*Difficulties.* An issue with this principle is that it allows for arbitrarily large inequalities in outcomes, and assumes a dubious distinction between ‘better-off’ and ‘worse-off’ (Leben, 2020). It thus is difficult, according to this criticism, to define what a harm is and what a benefit is. However, in the context of an STS this could perhaps be addressed through the contextual inputs from the social tier. Another issue with the principle of Non-Maleficence, however, is that just because harms are equally distributed does not necessarily mean that the outcome is fair. If there is an extremely unequal distribution of benefits, this still seems intuitively unfair; if the inequality of benefit distribution is large enough, it could plausibly become a harm. However, to counter this, as soon as it becomes a harm, the principle enforces that it should be equally distributed. Therefore perhaps it could be fair that benefits are unequally distributed only up to the point that it becomes harmful.

#### 4.2.2 EGALITARIANISM: EQUALITY OF OPPORTUNITY

**Egalitarian Equality of Opportunity** *Opportunities should be equally distributed in a way that ensures circumstances of birth or random choice are not held against individuals (Friedler et al., 2021).*

The goal of Equality of Opportunity is to ensure that negative attributes due to an individual’s circumstances of birth or random choice should not be held against them (Friedler et al., 2021). Yet, individuals should be still held accountable for their own actions. An individual’s well-being should thus be independent of their irrelevant attributes (Dwork et al., 2012).

*Operationalizing Equality of Opportunity.* Equality of Opportunity is a Deontological principle, and it would therefore be operationalized through the application of rules using a top-down, bottom-up, or hybrid architecture, supplemented by a definition of welfare. To implement Equality of Opportunity, Binns (2018) suggests considering whether each group is equally likely to be predicted a desirable outcome given the actual base rates for that group. Lee et al. (2021) argue it means that all opportunities should be equally open to all



applicants based on a relevant definition of merit. Welfare in this example would thus be access to opportunity.

*Difficulties.* However, in theory this can be fully satisfied even if only a minority segment of the population has realistic prospects of accessing the opportunity (Lee et al., 2021). As long as the opportunity is theoretically available, it is irrelevant as to whether it is practically accessible. It also may fail to address discrimination that may already lie within the data. Another argument against this is that it theoretically allows for a society in which some members end up in destitute conditions as long as they began in a place of equal opportunity (Fleurbaey, 2008).

#### 4.2.3 EGALITARIANISM: LUCK

**Egalitarian Luck** *No-one should end up worse off due to bad luck, and people should be given benefits as a result of their own choices (Lee et al., 2021).*

Descended from Equality of Opportunity, the aim of Luck Egalitarianism is understood as eliminating unchosen inequalities (Dworkin, 1981). It thus means no-one should end up worse off due to bad luck, but instead people should be given differentiated economic benefits as a result of their own choices (Lee et al., 2021).

*Operationalizing Luck.* There were no previous examples of Egalitarian Luck being operationalized in literature. As it is a Deontological principle, it could be implemented by applying rules in a top-down, bottom-up, or hybrid architecture, supplemented by a definition of welfare. In addition, the principle dictates that no-one should end up worse off due to bad luck; therefore, levels of luck must also be inputted into the decision mechanism.

*Difficulties.* A problem with this is that it is often difficult to separate out what is within an individual's genuine control. One needs to find in which circumstances, and to what extent, people should be held responsible for the unequal status they find themselves in (Binns, 2018). The ideal solution would then allow inequalities resulting from people's free choices and informed risk-taking, but disregard those which are the result of brute luck. They argue that the roles of notions like choice and desert should be considered: the choices made may deserve certain rewards and punishments, however where inequalities are the result of circumstances outside an individual's control, this should be corrected. Luck Egalitarianism thus only has a responsibility for creating advantages and disadvantages, and not for distributing them. However, Binns also points out that sometimes even inequalities which are the result of choice ought to be compensated, for example dependent caretakers.

#### 4.2.4 EGALITARIANISM: AUTONOMY

**Egalitarian Autonomy** *To obtain 'true' equality, a minimum level of autonomy must be obtained, with a minimum level of variety and quality of options, and a minimum decision-making competence (Fleurbaey, 2008).*

Equality of Autonomy has been proposed as including the full range of individual freedom (Lee et al., 2021). In order for their to be 'true' equality, a minimum level of autonomy must be attained, a minimum level of variety and quality of options should be offered, and there must exist a minimum decision-making competence (Fleurbaey, 2008). This is ar-

guably because in the absence of cost to others, it is desirable to give people more freedom and a greater array of choices in the future.

*Operationalizing Autonomy.* No examples of this principle being previously operationalized were found in literature. To operationalize the principle of Equality of Autonomy, it should be applied in the form of rules (as it is a Deontological principle) through a top-down, bottom-up, or hybrid architecture. As the principle denotes that minimum levels of autonomy should be attained, existing levels of autonomy must be implemented so that they can be fairly distributed in accordance with the principle.

*Difficulties.* However, when there is a significant asymmetry of power and information, autonomy in rational decision-makers fails as an ethical objective (Fleurbaey, 2008). In addition to this, this principle is only applicable to individual fairness, as it explicitly rests on the freedom of each individual. The weaknesses associated with individual fairness are thus inherited by the principle of autonomy, and casts doubt as to whether it would actually achieve a fair outcome.

### 4.3 Proportionalism

<p><b>Proportionalism</b> <i>The rights of individuals should be adjusted proportionally based on their contributions to production (Leben, 2020).</i></p>
------------------------------------------------------------------------------------------------------------------------------------------------------------

Proportionalism infers adjusting the rights of each person proportionally based on their contributions to production (Leben, 2020). This distribution should be managed based on factors that went into the process of production such as 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, and the amount of luck that went into those resources. The way proportional rights are conceived is commonly divided into two distinct approaches: Libertarian and Desert-Based.

*Operationalizing Proportionalism.* No examples of this principle being operationalized were found in literature. This could be done through the application of rules (as it is a Deontological principle) in a top-down, bottom-up, or hybrid architecture. In addition to a definition of welfare, levels of contribution should also be inputted to the decision mechanism as the principle states that rights should be adjusted in accordance with contributions to production.

*Difficulties.* An issue with Proportionalism is that there may be situations in which groups or individuals did not contribute that much to production, but should still be granted a distribution of rights. For example, a group that are unable to contribute due to disability should not be given less rights because of this. However, consideration of the influence of luck may mitigate this.

#### 4.3.1 PROPORTIONALISM: LIBERTARIAN

<p><b>Libertarian Proportionalism</b> <i>Each person is entitled to success rates in accordance with their total contribution at the time of consent (Leben, 2020).</i></p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Libertarian Proportionalism evaluates each person's total contribution at the time of consent (Leben, 2020). It sees each group as being entitled to success rates at least as fair as initial contributions. Inequality within that range is fair/acceptable (since it is pro-

portional to the original inequality), and any variance is unfair/unacceptable. Libertarian Proportionalists therefore only care about ensuring that inequality between groups does not exceed the pre-existing inequalities in the target trait. Libertarian ideals are understood as asserting the value of each person’s freedom insofar as there is no harm to anyone else, which naturally extends to the right to ownership and capital (Lee et al., 2021).

*Operationalizing Libertarianism.* Libertarian Proportionalism should be operationalized through the application of rules, as it is a Deontological principle. This could be done through a top-down, bottom-up, or hybrid architecture. In addition to a definition of welfare, levels of contribution would also need to be inputted as it is a form of Proportionalism.

Another approach to Libertarianism identified in literature applies the principle by allowing people to define ‘the good’ for themselves. For Etzioni and Etzioni (2016), Libertarians hold that each person should define the good and the values that are important, and the state should remain neutral. They suggest an ‘ethics bot’ which is able to apply to a variety of different pursuits and address moral choices in a similar way to AI targeted advertising, providing an interface between a person and other smart technologies.

*Difficulties.* A difficulty with this approach lies in the ability to gain consent. This can be a murky issue, especially in the context of group fairness. The extent to which a group of people can consent to something is unclear, and if this is not possible then it is difficult to analyse how much people have contributed and what benefits would be proportionate to this. Even if consent can be clearly obtained, there are also difficulties with Libertarianism in that in that it does not target pre-existing inequalities that may still be worth mitigating.

#### 4.3.2 PROPORTIONALISM: DESERT-BASED

**Desert-Based Proportionalism** *Rights are proportional to individual effort (Leben, 2020).*

Desert-based proportionalism, on the other hand, sees rights as proportional to individual effort (Leben, 2020). This is because prior prevalence of a trait in a population (which Libertarianism is based on) can be the result of unjust circumstances. Some literature understands desert as corresponding to virtue (Kagan, 1998).

*Operationalizing Desert-Based Proportionalism.* As it is a Deontological principle, Desert-Based Proportionalism could be implemented through the application of rules in a top-down, bottom-up, or hybrid architecture. A definition of welfare is necessary to understand what the principle is aiming for in its application. As it is a form of Proportionalism, contribution is also a required input. According to this principle, contribution is defined in terms of individual effort. Dwork (2012) implements Desert-Based Proportionalism by assigning each individual some distance in a metric space that evaluates desert, and then evaluates the fairness of the model by the average distance between individuals from each group in the metric space.

*Difficulties.* A weakness of this principle is that it is unclear as to what ‘unjust circumstances’ may be, which therefore makes it difficult to evaluate which traits should be mitigated for and which coordinate desert.

#### 4.4 Kantian

**Kantian** *The Categorical Imperative entails reasons for acting should be consistent with the assumption that all rational agents could engage in the same actions (Robbins & Wallace, 2007), and the Means-End Principle denotes that treating others as a means to an end is immoral (Abney, 2011).*

Kantian (2011) 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 (Kim et al., 2021). Kant’s *Categorical Imperative* means 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 (Robbins & Wallace, 2007; Abney, 2011; Kumar & Choudhury, 2022). All legitimate moral duties could be grounded in the Categorical Imperative (Wallach et al., 2008).

Derived from the Categorical Imperative is the *Means-End Principle*. This denotes that treating other people as a means to an end is immoral (Abney, 2011; Kumar & Choudhury, 2022). It would never be possible to universalize the treatment of another as a means to some end; doing so would contradict the Categorical Imperative. This is because all rational beings have intrinsic moral value, and the non-rational world has mere instrumental value. These two principles together arguably create an ideal world where society can act according to people’s maxim of will without affecting the welfare of others (Limarga et al., 2020).

*Operationalizing Kantian Deontology.* This principle could be operationalized through the application of rules (as it is Deontological) in a top-down, bottom-up, or hybrid architecture. This has been done in previous literature by Limarga et al. (2020) who implement the Categorical Imperative through the imposition of 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 its 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. This could be established through Berreby et al.’s (2017) technique to implement the Means-End Principle by defining the rule that an action is impermissible if it involves and impacts at least one person, but where at the same time that impact is not the aim of the action. An action is impermissible if it causes an event which involves at least one person, but where the event is not the aim of the action. Any other action is permissible.

*Difficulties.* An issue with the Categorical Imperative is that it is too permissive; it potentially permits intuitively bad things by allowing any action that can have a universalizable maxim (Abney, 2011). A common example of this is letting a murder 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 literally, it forbids any action in which a person affects another without their consent.

### 5. Teleological Ethical Principles

This section examines each Teleological principle identified in the review, including details as to how they have been previously operationalized and difficulties that may arise.

## 5.1 Virtue Ethics

**Virtue Ethics** *Ethicality stems from the inherent character of an individual, and not the rightness or wrongness of individual acts (Murukannaiah & Singh, 2020; Wallach et al., 2008; Brink, 2007; Kazim & Koshiyama, 2020; Robbins & Wallace, 2007; Yu et al., 2018; Abney, 2011).*

Virtue Ethics sees ethicality as stemming from the inherent character of an individual, and not the rightness or wrongness of individual acts; what counts is one’s moral character (Murukannaiah & Singh, 2020; Wallach et al., 2008; Brink, 2007; Kazim & Koshiyama, 2020; Robbins & Wallace, 2007; Yu et al., 2018; Abney, 2011). Right action is performed by someone with virtuous character, therefore in Virtue Ethics, one should not be asking what one ought to do, but rather what sort of person one should be (Anderson & Anderson, 2007; Rodriguez-Soto et al., 2022). The qualities one possesses should be primary and actions secondary. Virtues are described as dispositions to act in certain ways (Abney, 2011). Moral virtues can be learnt through habit and practice, which places virtue theory between the top-down explicit values advocated by a culture, and the bottom-up traits discovered or learned by an individual through practice. Instead of considering a specific situation or act, Virtue Ethics considers all actions of an individual’s life and examines whether these collectively constitute the actions of a virtuous person (Saltz et al., 2019).

*Operationalizing Virtue Ethics.* Implementing Virtue Ethics rests in developing virtues through a top-down, bottom-up, or hybrid architecture. Previous literature suggests that the stability of virtues (if one has a virtue, one can’t behave as if one doesn’t have it) entails that Virtue Ethics could be a useful way of imbuing machines with ethics (Wallach et al., 2008). Robbins and Wallace (2007) argue that to operationalize this principle, problems are solved ethically through the application of ‘virtuous’ characteristics. Vanhé and Borit (2022) suggest cultivating this through the education of designers of systems. Other works focus on implementing virtues directly into machines; according to Tolmeijer et al. (2021), inputs for implementing Virtue Ethics in machines would be properties of the agent, the decision criteria would be based on virtues, and the mechanism would be the instantiation of virtues. This is exemplified in Govindarajulu et al. (2017) through the use of computational formal logic to formalize emotions, represent traits, and establish a process of learning traits, to instantiate virtues. Greene et al. (2016) argue that a virtue based system would have to appreciate the whole variety of features in a given situation that would call for one action rather than another. Therefore, the development of virtuous characteristics is needed in order to implement Virtue Ethics into machines. This has been done in previously in literature by utilising computational formal logic to establish virtuous traits.

Virtue Ethics can also be used in conjunction with other approaches; Hagendorff (2020) argue that Deontological approaches should be augmented by combining it with Virtue Ethics through looking at values and character dispositions.

*Difficulties.* A problem with Virtue Ethics is that the holistic view it takes makes it more difficult to apply to individual situations or consider specific motivations (Saltz et al., 2019). Further challenges relate to conflicting virtues, and concretion of virtues (Tolmeijer et al., 2021). To judge whether a machine or human as virtuous is not possible by just observing one action or a series of actions that seem to imply the virtue – the reasons behind them 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 actually are. Additionally, the conception of virtues change greatly across time and culture, therefore those that are installed in machines now may lead to unfair outcomes in the future as virtues change.

## 5.2 Consequentialism

**Consequentialism** *The ethicality of an action is dependent on the effects of inequality on individuals and groups (Leben, 2020).*

In Consequentialist approaches, social justice is dependent on the *effects* of inequality on individuals and groups (Leben, 2020). Consequentialism is about identifying right actions, which are those that promote value (Brink, 2007; Yu et al., 2018; Cointe et al., 2016). The moral validity of an action can thus be judged only by taking its consequences into consideration (Saltz et al., 2019; Limarga et al., 2020; Rodriguez-Soto et al., 2022). A strength of this is that it can be used to evaluate decisions with complex outcomes where some benefit and some are harmed. It can thus explain many moral intuitions that trouble Deontological theories, as Consequentialists can say that the best outcome is the one in which the benefits outweigh the costs (Sinnott-Armstrong, 2021).

*Operationalizing Consequentialism.* Consequentialist principles can be operationalized by analysing the consequences of different actions, realised through a top-down, bottom-up, or hybrid architecture. This is exemplified in Limarga et al. (2020), who implement the principle by considering the consequences of different actions to make a proper judgement. This is taken further by Suikkanen (2017), who 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. (2021) argue that input for Consequentialist implementation would be the action (and its consequences), the decision criteria would be the comparative well-being, and the mechanism with which to achieve it would be the maximisation of utility (however, this applies to Utilitarianism specifically rather than consequentialism as a whole). For binary classification algorithms, Leben (2020) suggests that Consequentialism can be implemented by looking at how weights are assigned to each group outcome based on relative social cost.

*Difficulties.* However, assigning weights to each group outcome may be unrealistic to do for all protected groups (Leben, 2020). There might be high computational costs because Consequentialist systems would require a machine to represent all of the actions available to it (Greene et al., 2016). A related issue addressed lies in difficulties in estimating long-term or uncertain consequences and determining for whom consequences should be taken into account (Etzioni & Etzioni, 2017; Saltz et al., 2019). There are also moral constraints outside Consequentialism that prohibit certain actions even when they have the best outcomes, therefore rendering Consequentialist theories incomplete (Suikkanen, 2017). Another common criticism of Consequentialism concerns deciding what is valuable or intrinsically good: whether it is pleasure (hedonism), preference-satisfaction, perfection of one's essential capacities, or some list of disparate objective goods (e.g. knowledge, beauty, etc.) (Brink, 2007; Tolmeijer et al., 2021).

### 5.2.1 CONSEQUENTIALISM: UTILITARIANISM

**Utilitarianism** *An act is ethical if it maximizes the total expected utility across all who are affected (Kim et al., 2021; Lindner et al., 2019; Kazim & Koshiyama, 2020; Wallach et al., 2008; Murukannaiah & Singh, 2020; Mill, 1863).*

Utilitarianism is a Consequentialist theory that evaluates an act by its consequences; an act is ethical if and only if it maximizes the total net expected utility across all who are affected (Kim et al., 2021; Lindner et al., 2019; Kazim & Koshiyama, 2020; Wallach et al., 2008; Murukannaiah & Singh, 2020; Kumar & Choudhury, 2022; Rodriguez-Soto et al., 2022). The ultimate end is an existence exempt as far as possible from pain and as rich as possible in enjoyments (Mill, 1863). The greatest happiness principle thus states that actions are right in proportion to the happiness they promote, and wrong to the extent that they produce the reverse of happiness. Utility includes not just the pursuit of happiness but also the prevention or mitigation of unhappiness.

*Operationalizing Utilitarianism.* As Utilitarianism is a Consequentialist principle, it may be operationalized through analysing consequences of actions. In addition, this principle also requires a definition of utility and existing quantities of it. One example of this in previous literature is Leben (2020), who suggests implementing Utilitarianism to justify design choices for fairness metrics in binary classification algorithms. This can be done by constructing a function that models each potential distribution and its effects (a utility function/measure of happiness outcomes), and then running a selection procedure over aggregate utilities to maximise the sum. Similarly, Limarga et al. (2020) see Utilitarianism as denoting an act as right if and only if it maximises the good. To implement this, they assign some value to every action (the weight assigned to its worst consequence) which is used for final evaluation. Dehghani et al. (2008) also implement Utilitarianism in their MoralDM method by selecting the choice with the highest utility. Choudhury and Kumar (2022) suggest that based on the theory, AI agents could be trained to make judgements that deliver the greatest happiness to the greatest number of people.

*Difficulties.* A common problem with Utilitarianism is that it can lead to a smaller set of users being treated unfairly for the greater good (Ajmeri et al., 2020). It has also been argued that it is impossible to calculate the utility of every alternative course of action, and that the theory cannot readily account for the notion of rights and duties or moral distinctions between, for example, killing versus letting die (Abney, 2011). Another well-established issue with Utilitarianism relates to how utility can be quantified (for instance ‘higher’ and ‘lower’ pleasures) (Etzioni & Etzioni, 2017). To perhaps mitigate these issues, Utilitarianism could perhaps be seen as an additional necessary condition for an ethical action, rather than the sole ethical principle (Kim et al., 2021).

### 5.2.2 UTILITARIANISM: ACT UTILITARIANISM

**Act Utilitarianism** *The morally right action is the one that has the best overall consequences in terms of utility (Berreby et al., 2017; Anderson et al., 2004).*

Act Utilitarianism demands that one should assess the morality of an action directly in view of the principle of utility, which states that the morally right action is the one that has the best overall consequences (Berreby et al., 2017; Anderson et al., 2004). A machine

that acts utilising Act Utilitarianism can prompt considering alternative actions that might result in greater net good consequences, and to consider the effects of each of those actions on all those affected. It has been argued that nearly all Consequentialist machine ethics implementations utilize Act Utilitarianism (Tolmeijer et al., 2021).

*Operationalizing Act Utilitarianism.* The Consequentialist nature of this principle means it should be operationalized through analysing the consequences of actions. As it is a form of Utilitarianism, it also requires a definition of utility and existing quantities of it to determine the most appropriate solution. Berreby et al. (2017) implement Act Utilitarianism by determining an order of preferences between actions in the domain and then stating that an action is impermissible if there exists another action whose weight is greater.

*Difficulties.* A common Act Utilitarianism criticism is that one person can be sacrificed for the greater good, and also that it can conflict with a notion of justice or what people deserve (Anderson et al., 2004). This is because the rightness and wrongness of actions is determined entirely by their future consequences, whereas what people deserve is a result of past behaviour.

### 5.2.3 UTILITARIANISM: HEDONIC ACT UTILITARIANISM

**Hedonic Act Utilitarianism** *The morally right action is the one that derives the greatest net pleasure from all alternative actions (Anderson et al., 2004; Brink, 2007).*

Hedonic Act Utilitarianism entails computing the best action which derives the greatest net pleasure from all alternative actions (Anderson et al., 2004; Brink, 2007).

*Operationalizing Hedonic Act Utilitarianism.* As a Consequentialist principle, Hedonic Act Utilitarianism can be operationalized through analysing the consequences of actions. This principle is Hedonic therefore utility would be defined in terms of pleasure, and gathering levels of pleasure is crucial for obtaining a solution. In previous literature, Anderson et al. (2004) operationalize the principle by suggesting that as input, this principle requires the number of people affected, and for each person the intensity of pleasure/displeasure that will occur for each possible action. For each person, the algorithm then computes the product of intensity, the duration, and the probability to obtain the net pleasure for each person. This computation is performed for each alternative action.

*Difficulties.* A difficulty with this lies in that pleasure may not necessarily infer fairness. It is plausible that situations may arise which are really pleasurable for some, but greatly unfair for others – for instance, the humiliation of one person or group for the pleasure of everyone else.

### 5.2.4 UTILITARIANISM: RULE UTILITARIANISM

**Rule Utilitarianism** *The morally right action is assessed through understanding whether a (set of) moral rule(s) will lead to the best overall consequences, assuming all or most agents follow it (Berreby et al., 2017).*

Rule Utilitarianism involves morally assessing an action by first appraising moral rules on the basis of the principle of utility – deciding whether a (set of) moral rule(s) will lead to the best overall consequences, assuming all or at least most agents follow it (Berreby et al., 2017). For example, one such rule may be ‘do not steal’. The second step consists in the



appraisal of particular actions in light of what was justified in the first step. An action is permissible only if the action is sanctioned by a rule that upholds the principle of utility, whether or not the action itself adheres to the principle of utility. Therefore, contrasting Act Utilitarianism, the issue is not which action produces the greatest utility, but which moral rule does.

*Operationalizing Rule Utilitarianism.* This principle is Consequentialist, and therefore may be operationalized through analysing the consequences of actions, and requires a definition of utility to do so. Berreby et al. (2017) implement this principle by using a predicate which compounds all the effective weights of the actions that belong to a particular rule, then summing up those weights via a predicate. An action is then seen as impermissible if there is an instance of a rule that is overall harmful, i.e. an instance of a rule whose bad consequences outweigh its good ones, considering all of its instantiations.

*Difficulties.* A common problem with Rule Utilitarianism is that 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.

### 5.3 Consequentialism: Prioritarianism/Maximin

**Prioritarianism/Maximin** *The minimum utility should be maximised by improving the worst-case experience in a society (Ajmeri et al., 2020).*

The Maximin principle rests in maximising the minimum utility by seeking to improve the worst-case experience in a society; guaranteeing a higher than worst-case minimum utility to each individual (Ajmeri et al., 2020). It shifts the focus towards the improvement of the well-being of those who are worst-off (Lee et al., 2021). The difference principle states that economic and social inequalities can only be justified if they benefit the most disadvantaged members of society (Rawls, 1967). Prioritarianism/Maximum thus focuses on improving the worst-case situations.

*Operationalizing Prioritarianism.* Prioritarianism is Consequentialist, and therefore can be operationalized by analysing consequences. It also requires a definition of utility to understand how utility can be distributed. This principle has been implemented in literature, for example by Ajmeri et al. (2020) who propose a technical agent that aims to improve the minimum experience/worst-case outcome for any user. Leben (2020) argues that this principle can be used to justify design choices of fairness metrics for binary classification algorithms by constructing a function that models each potential distribution and its effects. Then, one should run a selection procedure over aggregate utilities, maximising the minimum. Diana et al. (2021) present algorithms using the ‘minimax’ framework in which fairness is measured by worst-case outcomes across all groups, rather than differences between group outcomes. Their goal is thus to minimise the maximum loss across all groups, rather than equalising group losses, to make sure that the worst-off group is as well-off as possible. Another example is Sun et al. (2021), who use the Maximin principle in the chores allocation problem by minimising the maximum cost of an allocation over all allocations.

*Difficulties.* One issue with this is that although the aggregate utility may be increased, it does not necessarily mitigate the effects of discrimination (Sun et al., 2021). It still allows for disparities between groups. Therefore, the most privileged group may still remain

much more privileged than the least privileged group, despite the overall experience being improved.

### 5.3.1 CONSEQUENTIALISM: ENVY-FREENESS

**Envy-Freeness** *The ethical action should be the one in which no agent envies another agent (Sun et al., 2021).*

In an Envy-Free allocation, no agent envies another agent (Sun et al., 2021). 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.

*Operationalizing Envy-Freeness.* This is a Consequentialist principle and can therefore be operationalized by analysing consequences. In doing so, levels of envy must be inputted. Boehmer and Niedermeier (2021) argue that an assignment of resources to agents is Envy-Free if no agent prefers another agent’s bundle (of resources) to their own.

*Difficulties.* One issue is the argument that what is important isn’t a relative condition to other people, but whether people have enough to have satisfactory life prospects (Lee et al., 2021). In addition, it may be hard to get accurate measures of envy, as it is a subjective entity that stakeholders may not always be open about.

Another issue is that the existence of an Envy-Free allocation can’t be guaranteed when items to be assigned are indivisible, for example chores that need to be assigned to multiple agents (Sun et al., 2021). This has led to relaxations of Envy-Freeness such as Envy-Free up to one item (one agent may be jealous of another, but by removing one chore from the bundle of the envious agent, envy can be eliminated), and Envy-Free up to any item (envy can be eliminated by removing any positive-cost chore from the bundle of the envious agent).

### 5.3.2 CONSEQUENTIALISM: DOCTRINE OF DOUBLE EFFECT

**Doctrine of Double Effect** *Deliberately inflicting harm is wrong, even if it leads to good (Deng, 2015).*

The Doctrine of Double Effect (DDE) suggests that deliberately inflicting harm is wrong, even if it leads to good (Deng, 2015). On the other hand, inflicting harm might be acceptable if it is not deliberate, but simply a consequence of doing good. For this principle, an action is permissible if the action itself is morally good or neutral, some positive consequence is intended, no negative consequence is a means to the goal, and the positive consequences sufficiently outweigh the negative ones (Lindner et al., 2019). It has been explained as allowing actions that have both positive and negative effects in situations where both positive and negative effects appear to be unavoidable (Govindarajulu & Bringsjord, 2017).

*Operationalizing the Doctrine of Double Effect.* This principle can be operationalized through analysing the consequences of actions. Govindarajulu and Bringsjord (2017) automate DDE, and also the stronger version of the Doctrine of Triple Effect, using formal logic to operationalize the principle. They use the framework in two different modes: to build DDE-compliant autonomous systems from scratch, or to verify that a given AI system is DDE-compliant. Another approach by Berreby et al. (2017) implements this principle by

having rules that proscribe an action if it is intrinsically bad, if it causes a bad effect which leads to a good effect, and if its overall effects are bad.

*Difficulties.* An issue identified with DDE is that it still allows bad actions to happen as long as they are not intended, which may have some morally dubious outcomes.

### 5.3.3 CONSEQUENTIALISM: DISPARATE IMPACT DOCTRINE

**Disparate Impact Doctrine** *Any group should have equal or proportional representation in the outcome (Patel et al., 2020).*

The Disparate Impact Doctrine has been suggested for use as a notion of group fairness (Patel et al., 2020). It posits that any group must have approximately equal or proportional representation in the solution provided by the algorithm. The concept of ‘disparate mistreatment’ has also been suggested, which considers differences in false positive rates between groups (Binns, 2018). It thus emphasises the importance of ensuring impact is proportionally distributed amongst the relevant groups.

*Operationalizing the Disparate Impact Doctrine.* This principle can be operationalized through analysing the consequences of actions, as it is Consequentialist. For example, by using the contextual inputs to evaluate an action that leads to proportional or equal representation of groups in the solution. Disparate mistreatment would ensure that the existence of people being treated wrongly is equal or proportional.

*Difficulties.* An issue with this, which is a common complaint against group fairness, is that it may lead to individuals being unfairly treated in favour of the group.

### 5.3.4 CONSEQUENTIALISM: DO NO HARM

**Do No Harm** *Inflicting harm in any capacity is wrong (Lindner et al., 2019).*

This principle enforces looking at the harm caused by an action, stating that no harm should be inflicted. Any action that causes harm would thus be unethical.

*Operationalizing Do No Harm.* The principle of Do No Harm is Consequentialist, and therefore can be operationalized through analysing the consequences of actions. Lindner (2019) implements this by stating that a technical agent may not perform an action which causes any harm. Dennis et al. (2016) suggests an example of Do No Harm being violated by a specific action is by moving ten metres to the left when an aircraft is on the ground, therefore ethically constraining the aircraft from performing this action in this circumstance.

*Difficulties.* 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.

### 5.3.5 CONSEQUENTIALISM: DO NO INSTRUMENTAL HARM

**Do No Instrumental Harm** *Harm is allowed as a side effect, but not as a means to a goal (Lindner et al., 2019).*

This principle allows for harm as a side effect, but not as a means to a technical agent’s goals (Lindner et al., 2019).

*Operationalizing Do No Instrumental Harm.* This principle could be implemented in largely the same way as the principle of Do No Harm, except that it would allow for harm to be caused as a side effect.

*Difficulties.* This may help situations in which causing harm is inevitable, although it does still allow that some harm is acceptable, which may lead to certain groups or individuals being treated unfairly.

## 5.4 Other Principles

In addition to the principles mapped out here, there are other principles mentioned in literature. These have not been included in the taxonomy for a variety of reasons, as shall be explained here.

*Egoism* is acting to reach the greatest outcome possible for one’s self, irrespective to others (Robbins & Wallace, 2007; Kumar & Choudhury, 2022). This principle was 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 fairness would be an ethical goal for them. This is because fairness is aimed at the well-being of others as well as the self, whereas egoism is solely self-centred.

*Pluralism* states that there is no one approach that is best (Robbins & Wallace, 2007). Using context and various reasoning techniques to choose between principles is appropriate. Tolmeijer et al. (2021) also advocate for further research according to this approach, applying multi-theory models where machines can interchangeably apply different theories depending on the type of situation. They argue that human morality is complex and cannot be captured by one single classical ethical theory. This principle was not included in this taxonomy as it is not itself a principle in that it does not guide a particular course of action. However, it is a useful approach to have for ethics, and perhaps could be relevant to help developers to understand how to utilise ethical principles.

*Particularism* emphasizes that there is no unique source of normative value, nor is there a single, universally applicable procedure for moral assessment (Tolmeijer et al., 2021). 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. The authors argue that inputs for Particularism would include the situation (context, features, intentions, and consequences), and the decision criteria rests on rules of thumb and precedent, as all situations are unique. The mechanism to decide upon an action lies in how much it fits with rules or precedent. Some challenges they identify are that there are no unique and universal logic, thus each situation needs a unique assessment. The lack of a universal logic for it is part of the reason for not including it in the taxonomy: it does not give clear guidance. However, Particularism is perhaps relevant to the ways in which ethical principles can be operationalized, as it emphasizes the inclusion of context in the moral reasoning process. It is not in itself an operationalizable ethical principle, but perhaps more of a meta-principle that can be used in the application of other ethical principles.

*The ethic of care and responsibility* relates to considering your feelings of interconnectedness with others (Robbins & Wallace, 2007; Martin, 2022). 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 having the goal of fairness, however in itself is not a principle in that it denotes a certain action. It may be related to the Kantian Means-End principle, however this has not been included as an individual principle in the taxonomy.

*Other cultures.* Lastly, there is a wide variety of principles proposed in cultures outside of the history of Western ethics. Moral frameworks have been established in societies across the world, including Confucian, Shinto, and Hindu thought as well as religious frameworks like Judaism, Christianity, and Islam (Hagerty & Rubinov, 2019). They argue that there is a multitude of moral frameworks across cultures with significant variation within these frameworks. For the authors, ethics and culture are inseparable and to understand one you must look at the other. Therefore, they argue that ethics must be considered within its cultural context. The reason these principles were not included in the taxonomy was not because they are unimportant, but because they would require a whole taxonomy 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, with the goal of forming a taxonomy of such principles. This is crucial to help developers to build cross-cultural ethical technology.

## 6. Conclusions and Directions

In order to better address the pursuit of fairness in AI, research must be human-centred (Dignum & Dignum, 2020). Shifting the perspective to macro ethics of STS in which the governance of systems is examined is crucial to better achieve this (Chopra & Singh, 2018). In the governance of STS, stakeholders attempt to align norms with their values. However, dilemmas in decision making can arise when stakeholders have different preferences (Murukannaiah et al., 2020). 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 (McLaren, 2003; Lindner et al., 2019).

### 6.1 Key takeaways and Directions

Based on our review, we identify key takeaways that practitioners should know in order to implement ethical principles. We envision that the taxonomy we develop and the key takeaways we identify will help to enable the operationalization of ethical principles in the governance of STS. Figure 3 summarise the takeaways.

**Clarifying the Architecture.** In the implementation of ethical principles, designers must decide as to whether principles are being implemented through a top-down, bottom-up, or hybrid approach to the architecture (Wallach et al., 2008).

- *Direction.* Opportunities in this area lie in further researching bottom-up approaches, as well as formal ways of discerning the circumstances in which each of the different architectures is appropriate.

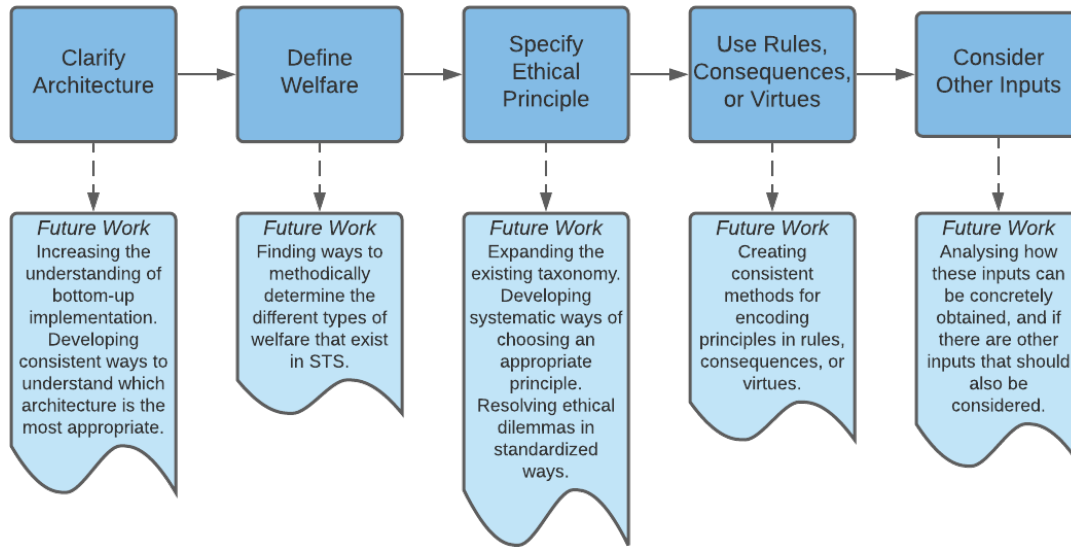


Figure 3: Key Takeaways and Future Research to Operationalize Ethical Principles in STS

**Defining Welfare.** Having a clear definition of welfare to specify what is good or valuable is necessary in the operationalization of ethical principles in order to understand what constitutes a satisfactory outcome (Fleurbaey, 2008).

- *Direction.* To properly understand how welfare should be defined, more work is needed in the different types of welfare that exist in relation to STS and how this can be methodically determined.

**Specifying Ethical Principles.** Researchers should be specific about which ethical principles they are using in their research (Binns, 2018). A broad array of ethical principles were found from AI and computer science literature, with 25 nodes on the taxonomy tree (Figure 2). There are likely more that exist but they were not identified in this review. 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 (Leben, 2020). In addition to this, researchers should promote the use of lesser-utilized principles to avoid a monopoly of commonly known principles such as Utilitarianism.

- *Direction.* This therefore presents opportunities in the incorporation of principles that were not identified in this review (especially those outside of the Western doctrine), and emphasis on future work being explicit on the inclusion of ethical principles.
- *Direction.* In addition to this, the development of consistent methodologies to choose which ethical principle is appropriate would be beneficial to practitioners. This is because the abstraction of normative ethical theories entails that their suitability can depend on a variety of factors including personal preference, norms, values, and context. Method-

ologies to help identify ethical principles appropriate to specific scenarios would therefore be useful.

**Rules, Consequences, or Virtues.** There are three key directions to implement the ethical principles identified. The first is by aligning actions to certain rules (Greene et al., 2016). The second is by choosing actions based on the (potential) consequences that they produce, and which consequence is best according to the chosen principle (Suikkanen, 2017). The third is through the instantiation of virtues in machines; right action is thus produced by virtuous machines (Govindarajulu et al., 2019).

- *Direction.* Looking forward, the encoding of principles in either rules, consequences, or virtues will require a great deal of further research to develop systematic methods in the context of governing STS.

**Principle Dependent Inputs.** There is a selection of inputs that vary from principle to principle as to whether or not they are necessary. This includes the consideration of luck, autonomy, contribution, virtue, utility, and envy.

- *Direction.* These inputs run the risk of being somewhat abstract, and perhaps difficult to infer. Further research is therefore needed as to how they can be obtained, and if there are even more inputs that were not identified here.

## References

- Abney, K. (2011). *Robots, Ethical Theory, and Metaethics: A Guide for the Perplexed*, pp. 35–52. MIT Press, Cambridge.
- Ajmeri, N., Guo, H., Murukannaiah, P. K., & Singh, M. P. (2020). Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 16–24, Auckland. IFAAMAS.
- Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), 15.
- Anderson, M., & Anderson, S. L. (2014). GenEth: A general ethical dilemma analyzer. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 1, pp. 253–261, Québec. Association for the Advancement of Artificial Intelligence.
- Anderson, M., Anderson, S. L., & Armen, C. (2004). Towards machine ethics. In *AAAI-04 Workshop on Agent Organizations: Theory and Practice*, pp. 1–7, San Jose. AAAI.
- Berreby, F., Bourgne, G., & Ganascia, J.-G. (2017). A declarative modular framework for representing and applying ethical principles. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 96–104, São Paulo. International Foundation for Autonomous Agents and Multiagent Systems.
- Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In Friedler, S., & Wilson, C. (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81 of *Proceedings of Machine Learning Research*, pp. 149–159, New York. PMLR.
- Bishr, A. B. B. (2018). AI ethics principles & guidelines. Smart Dubai.

- Boehmer, N., & Niedermeier, R. (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)*, pp. 1–5, London. International Foundation for Autonomous Agents and Multiagent Systems.
- Brink, D. (2007). Some forms and limits of consequentialism. *The Oxford Handbook of Ethical Theory*, 1(1), 381–423.
- Britannica (2021). Teleological ethics. <https://www.britannica.com/topic/teleological-ethics>. Accessed: 2021-09-23.
- Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *AI Magazine*, 38(2), 22–34.
- Canca, C. (2020). Operationalizing AI ethics principles. *Communications of the ACM*, 63(12), 18–21.
- Cheng, L., Varshney, K., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *JAIR*, 71, 1137–1181.
- Chopra, A., & Singh, M. (2018). Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 48–53, New Orleans. Association for Computing Machinery.
- Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, 99(4), 906–944.
- Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical judgment of agents’ behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1106–1114, Singapore. IFAAMAS.
- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4831–4835, Honolulu. AAAI.
- Cummiskey, D. (1990). Kantian consequentialism. *Ethics*, 100(3), 586–615.
- Dastani, M., & Yazdanpanah, V. (2022). Responsibility of AI Systems. *AI & SOCIETY*, 1(1435-5655).
- Dechesne, F., Di Tosto, G., Dignum, V., & Dignum, F. (2013). No smoking here: Values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1), 79–107.
- Dehghani, M., Tomai, E., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. *Machine Ethics*, 3, 1280–1286.
- Deng, B. (2015). Machine ethics: The robot’s dilemma. *Nature*, 523, 24–26.
- Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77, 1–14.
- Diana, E., Gill, W., Kearns, M., Kenthapadi, K., & Roth, A. (2021). Convergent algorithms for (relaxed) minimax fairness. *CoRR*, abs/2011.03108, 1–22.
- Dignum, V. (2019). *Ethical Decision-Making*, pp. 35–46. Springer, Cham.



- Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M. S., Lopez-Sanchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman, M., van Steenberghe, M., Tedeschi, S., van der Toree, L., Villata, S., & de Wildt, T. (2018). Ethics by design: Necessity or curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 60–66, New York, NY, USA. Association for Computing Machinery.
- Dignum, V., & Dignum, F. (2020). Agents are dead. long live agents!. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1701–1705, Auckland. International Foundation for Autonomous Agents and Multiagent Systems (AAMAS).
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226, Cambridge. ACM.
- Dworkin, R. (1981). What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs*, 10(3), 185–246.
- Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, 18, 149–156.
- Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21, 403–418.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (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*, pp. 1–39, Cambridge. Berkman Klein Center.
- Fleurbaey, M. (2008). *Fairness, Responsibility, and Welfare*. Oxford University Press, Oxford.
- Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 1. <https://hdr.mitpress.mit.edu/pub/10jsh9d1>.
- Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM (CACM)*, 64(4), 136–143.
- Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 4722–4730, Melbourne. IJCAI.
- Govindarajulu, N. S., Bringsjord, S., Ghosh, R., & Sarathy, V. (2019). Toward the engineering of virtuous machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 29–35, Honolulu, USA. Association for Computing Machinery.
- Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. (2016). Embedding ethical principles in collective decision support systems. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4147–4151, Snowbird. AAAI Press.

- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30, 99–120.
- Hagerty, A., & Rubinov, I. (2019). Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. *CoRR*, abs/1907.07892, 1–27.
- Honarvar, A. R., & Ghasem-Aghaee, N. (2009). An artificial neural network approach for creating an ethical artificial agent. In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation - (CIRA)*, pp. 290–295, Daejeon. IEEE.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kagan, S. (1998). *Equality and Desert*, chap. 30, pp. 298–314. Oxford University Press, Oxford.
- Kant, I. (2011). *Immanuel Kant: Groundwork of the Metaphysics of Morals: A German-English edition*. The Cambridge Kant German-English Edition. Cambridge University Press, Cambridge.
- Kazim, E., & Koshiyama, A. (2020). A high-level overview of AI ethics. *SSRN*, 1(1), 1–18.
- Khan, A. A., Badshah, S., Liang, P., Khan, B., Waseem, M., Niazi, M., & Akbar, M. A. (2021). Ethics of AI: A systematic literature review of principles and challenges. *CoRR*, abs/2109.07906.
- Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment. *JAIR*, 70, 871–890.
- Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Tech. rep., Keele University and Durham University Joint Report.
- Kökciyan, N., Yaglikci, N., & Yolum, P. (2017). An argumentation approach for resolving privacy disputes in online social networks. *ACM Trans. Internet Technol.*, 17(3), 1–22.
- Kökciyan, N., & Yolum, P. (2020). Turp: Managing trust for regulating privacy in internet of things. *IEEE Internet Computing*, 24(6), 9–16.
- Kumar, S., & Choudhury, S. (2022). Normative ethics, human rights, and artificial intelligence. *AI & Ethics*, 2, 1–10.
- Leben, D. (2020). Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 86–92, New York. Association for Computing Machinery.
- Lee, M., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics*, 1(1), 529–544.
- Liao, B., Slavkovik, M., & van der Torre, L. (2019). Building Jiminy Cricket: An architecture for moral agreements among stakeholders. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 147–153, Honolulu. ACM.

- Limarga, R., Pagnucco, M., Song, Y., & Nayak, A. (2020). Non-monotonic reasoning for machine ethics with situation calculus. In *AI 2020: Advances in Artificial Intelligence*, pp. 203–215, Canberra. Springer International Publishing.
- Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence*, 175(5), 942–949. Special Review Issue.
- Lindner, F., Mattmüller, R., & Nebel, B. (2019). Moral permissibility of action plans. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 7635–7642.
- Manjarrés, Á., Fernández-Aller, C., López-Sánchez, M., Rodríguez-Aguilar, J. A., & Castañer, M. S. (2021). Artificial intelligence for a fair, just, and equitable world. *IEEE Technology and Society Magazine*, 40(1), 19–24.
- Martin, K. (2022). *Ethics of Care as Moral Grounding for AI*, pp. 1–6. Auerbach Publications.
- McLaren, B. M. (2003). Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence*, 150(1), 145–181. AI and Law.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
- Mill, J. S. (1863). *Utilitarianism*. Longmans, Green and Company.
- Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, 31, 239–256.
- Morris-Martin, A., De Vos, M., & Padget, J. (2019). Norm emergence in multiagent systems: A viewpoint paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 33(6), 706–749.
- Murukannaiah, P. K., Ajmeri, N., Jonker, C. M., & Singh, M. P. (2020). New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1706–1710, Auckland. IFAAMAS. Blue Sky Ideas Track.
- Murukannaiah, P. K., & Singh, M. P. (2020). From machine ethics to Internet ethics: Broadening the horizon. *IEEE Internet Computing*, 24(3), 51–57.
- Patel, D., Khan, A., & Louis, A. (2020). Group fairness for knapsack problems. *CoRR*, abs/2006.07832, 1–36.
- Pedamkar, P. (2021). Intelligent agent in AI. <https://www.educba.com/intelligent-agent-in-ai/>. Accessed:2021-12-21.
- Rădulescu, R., Mannion, P., Roijers, D. M., & Nowé, A. (2019). Multi-objective multi-agent decision making: A utility-based analysis and survey. *CoRR*, abs/1909.02964, 1–48.
- Rawls, J. (1958). Justice as fairness. *The Philosophical Review*, 67(2), 164–194.
- Rawls, J. (1967). Distributive justice. *Philosophy, Politics and Society*, 1, 58–82.
- Rawls, J. (1985). Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs*, 14(3), 223–251.

- Robbins, R., & Wallace, W. (2007). Decision support for ethical problem solving: A multi-agent approach. *Decision Support Systems*, 43(4), 1571–1587. Special Issue Clusters.
- Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2022). Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1), 9.
- Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.*, 19(4), 1–26.
- Schwartz, S. H. (2012). An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1), 2307–0919.
- Sen, A. (1992). *Inequality reexamined*. Clarendon Press, Oxford.
- Singh, M. P. (2013). Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1), 21:1–21:23.
- Sinnott-Armstrong, W. (2021). Consequentialism. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 edition). Metaphysics Research Lab, Stanford University, Stanford.
- SOAS (2021). Unit 1 introduction to ethics. [https://www.soas.ac.uk/cedep-demos/000\\_P563\\_EED\\_K3736-Demo/unit1/page\\_17.htm#](https://www.soas.ac.uk/cedep-demos/000_P563_EED_K3736-Demo/unit1/page_17.htm#). Accessed: 2021-09-23.
- Suikkanen, J. (2017). Consequentialism, constraints, and good-relative-to: A reply to mark schroeder. *Journal of Ethics and Social Philosophy*, 3(1), 1–9.
- Sun, A., Chen, B., & Doan, X. V. (2021). Connections between fairness criteria and efficiency for allocating indivisible chores. *CoRR*, *abs/2101.07435*, 1–32.
- Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2021). Implementations in machine ethics: A survey. *ACM Comput. Surv.*, 53(6).
- Vanhée, L., & Borit, M. (2022). Viewpoint: Ethical by designer - how to grow ethical designers of artificial intelligence. *JAIR*, 73, 619–631.
- Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI and Society*, 22(4), 565–582.
- Woodgate, J., & Ajmeri, N. (2022). Macro ethics for governing equitable sociotechnical systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1824–1828, Online. IFAAMAS. Blue Sky Ideas Track.
- Yazdanpanah, V., Gerding, E., Stein, S., Dastani, M., Jonker, C. M., & Norman, T. (2021). Responsibility research for trustworthy autonomous systems. In *20th International Conference on Autonomous Agents and Multiagent Systems (03/05/21 - 07/05/21)*, pp. 57–62.
- Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI*, pp. 5527–5533, Stockholm. IJCAI.
- Zhu, L., Xu, X., Lu, Q., Governatori, G., & Whittle, J. (2022). *AI and Ethics—Operationalizing Responsible AI*, pp. 15–33. Springer International Publishing, Cham.

## Appendix A. Overview of Method

Figure 4 visualises the method used in order 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 and Table 2). 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., Leben, 2020). These works were useful for the identification of principles ( $Q_P$ ). Other research took established principles and implemented them, which helped to answer  $Q_O$  (e.g., Sun et al., 2021). Some works had a mixture of both identification and implementation (e.g., Kim et al., 2021). 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.

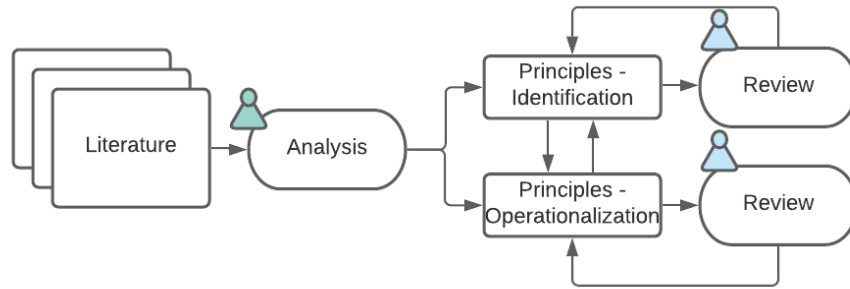


Figure 4: Methodology to Extract Principle Identification and Operationalization from Literature

## Appendix B. 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 to 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 international authorship, but it is recognized 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 off of

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, it is ensured that relevant citations are followed and those terms are included.

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 (2007) 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 the 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 3: Inclusion and Exclusion Criteria

<b>Inclusion</b>	<b>Exclusion</b>
Published works found in: AIES, FAccT, AAAI, IJCAI, (J)AAMAS, TAAS, TIST, JAIR, AIJ, Nature, Science	Works about meta-ethics or applied ethics outside of computer science
Individual and/or group fairness	Studies about specific ML methodology
Multiple-user social dilemmas	Non-social dilemmas
Normative ethics and multiple-user AI and/or MAS	Non-ethical studies of multiple-user AI and/or MAS
Normative ethics and STS	Non-ethical studies of STS
Normative ethical principles and AI	AI keystones
Bias when related to ethical principles	Studies about bias without reference to ethical principles