Finally, it should be noted that humans, as moral agents, never reason using only one single theory, but will switch between different theories, or adaptations thereof, depending on many circumstances. This is not only because humans are not the pure rational agents that economic theories would like us to believe, but also because strict following of any ethical theory leads to unwanted results. For example, a strict application of utilitarian theory would justify making one (or a few) miserable in service of the rest, because all what matters is the net global happiness. However, few of us would choose for actions through which one person is explicitly and consciously sacrificed for the benefit of many. In fact, utilitarianism ignores the unjust distribution of (good) consequences, and deontologic models allow no exceptions to moral rules, given their assumption that 'the law' is always ethical. This would mean that an AI system should be given representations of different ethical theories, and the capability to choose between these.

5.3 Approaches to Ethical Reasoning by AI

The examples and considerations in the previous section show only some of the many complexities of ethical reasoning by AI systems. In this section, we will discuss current approaches and reflect on their consequences.

Existing approaches to the implementation of ethical reasoning can be divided into three main types [132, 12]:

- **Top-down approaches**, which infer individual decisions from general rules. The aim is to implement a given ethical theory, such as those discussed in Chapter 3, within a computational framework, and apply it to a particular case.
- **Bottom-up approaches**, which infer general rules from individual cases. The aim is to provide the agent with sufficient observations of what others have done in similar situations and the means to aggregate these observations into a decision about what is ethically acceptable.
- **Hybrid approaches**, which combine elements from bottom-up and top-down approaches in support of a careful moral reflection which is considered essential for ethical decision-making [102].

Taking a top-down approach the premise is that it is possible to implement some kind of $eth(a, c)$ function. This is easily said but difficult to do, as we saw in the previous section. To calculate this function, first we need to determine which ethical value we are maximising for. Is that e.g. fairness, human dignity or maybe trustworthiness? It is easy to see that maximising for fairness may provide different results than maximising for human dignity, or for safety, or for privacy, depending also on how these values are implemented: cf. Section 4.4 for an example. And of course, as we have seen in Chapter 3 taking a utilitarian stance results in very different choices than taking a

deontological or virtues stance. Who determines which are the choices and how to implement $eth(a, c)$? Responsible AI is about answering this question, and answering for the consequences of these decisions. This requires a higher level of reflection and abstraction than that needed to decide on the implementation. In Section 5.3.1 we present some current approaches to this issue.

On the other hand, taking a bottom-up approach means that, in a way, we are equating social acceptability with ethical acceptance. That is, we assume that what the other agents are doing is the ethical thing to do. This would mean that $eth(a, c)$ would be dynamically built from observations of what others are doing, and the evaluation of the perceived results of those actions. Even though this is actually the process by which children typically learn ethical behaviour, it is one that may lead to large differences depending on the examples that are made available to the system to learn from. As shown in the recent analysis of the large online experiment with the Moral Machine, cultural and contextual factors lead to very different choices [10]. Again here, a higher level of reflection is needed: from whom is the system going to learn, and who decides that? And also, which data is going to be collected about the behaviour of the crowd and how is that data to be aggregated? In Section 5.3.2 we discuss this issue further.

Finally, hybrid approaches combine characteristics from both approaches in an attempt to approximate human ethical reasoning. These approaches typically provide some *a priori* information about legal behaviour and enable agents to decide on action by observing what others do within the limits of what is allowed by the rules. In Section 5.3.3 we discuss this approach further.

5.3.1 Top-Down Approaches

A top-down approach to modelling ethical reasoning assumes a given ethical theory (or possibly, a set thereof), and describes what the agent ought to do in a specific situation, according to that theory. These models formally define rules, obligations and rights that guide the agent in its decision-making. Top-down approaches are often an extension of work on normative reasoning, and in many cases are based on Belief-Desire-Intention architectures. Normative systems, such as the ones we have developed in previous work [39], take a deontological approach, assuming that following existing laws and social norms guarantees 'good' decisions. Deontological approaches have been extensively formalised using e.g. Deontic Logics.

Top-down approaches differ in the ethical theory that is chosen. So, maximising models roughly follow a Utilitarian view, by taking into account the satisfaction of a given value as the basis for the decision ('the best for the most'), whereas models that follow a Deontological view will evaluate the 'goodness' of the actions themselves. In a recent paper, [111], the authors

propose to specify the moral values associated with behaviour norms as an additional decision criterion beyond those regarding norm representation and associated costs. Another approach is to endow an agent with an internal representation of values and ethics to judge the ethical aspects of its own behaviour and that of other agents in a multi-agent system [29].

Top-down approaches assume that AI systems are able to explicitly reason about the ethical impact of their actions. Such systems should meet the following requirements:

- Representation languages rich enough to link domain knowledge and agent actions to the values and norms identified;
- Planning mechanisms appropriate to the practical reasoning prescribed by the theory; and
- Deliberation capabilities to decide whether the situation is indeed an ethical one.

Several computational architectures meet these requirements, but research is needed to determine whether this is in fact a responsible approach to ethical behaviour. In this section, we merely aim to provide a sketch of the possibilities rather than a full account of architectural and implementation characteristics.

Reflection on the top-down approach

Ethical theories provide an abstract account of the motives, questions and aims of moral reasoning. Section 3.5 presented several design options concerning who is responsible for a decision, and how decisions are dependent on the relative priority of different moral and societal values. However, and despite sincere attempts at top-down models of ethical reasoning, for its practical application more is needed, namely the understanding of which moral and societal values should be at the basis of deliberation in different situations, and how the agent should deliberate. For example, Consequentialistic approaches aim at 'the best for the most' but one needs to understand societal values in order to determine what counts as the 'best'. In fact, depending on the situation, this can be e.g. wealth, health, sustainability or even a combination of values.

Top-down approaches impose a system of ethics on the agent. These approaches implicitly assume a similarity between ethics and the law, i.e. that a set of rules is adequate and sufficient as a guide for ethical behaviour. However, these are not identical. Typically, the law tells us what we are prohibited from doing and what we are required to do. The law tells us what are the rules of the game, but gives no support on how to best win the game, whereas ethics tells us how to play a 'good' game for all.

Moreover, something may be legal but we may consider it unacceptable. And we may consider something right but it may not be legal. So, by equating

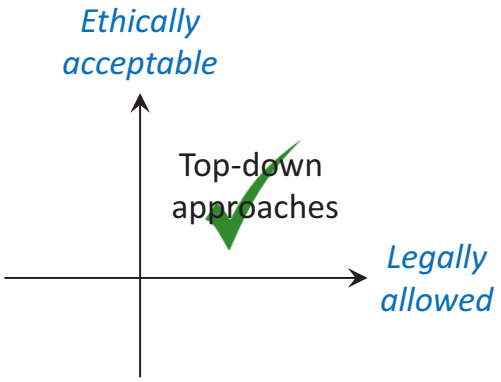

Figure 5.1: Top-down approaches assume alignment between law and ethics

what is *legally allowed*, as defined by a set of rules, with what is *ethically acceptable*, we are possibly disregarding many other possible attitudes to what is the ethical thing to do in a situation (cf. Figure 5.1).

Nevertheless, given that AI systems are artefacts built for a given purpose, it may be correct to demand that these artefacts to stay within the realm of what is both legal and ethical, and do not learn other options by themselves. That is, AI systems should be seen as incorporating soft ethics that see ethics as post-compliant to an existing regulatory system, and used to decide on what ought and ought not to be done over and above the existing regulation [55].

5.3.2 Bottom-Up Approaches

Bottom-up approaches assume that ethical behaviour is learned from observation of the behaviour of others. According to Malle a morally competent robot should be equipped with a mechanism that allows for *'constant learning and improvement'* [85, p.11]. He states that robots need to learn norms and morality, like little children do, in order to become ethically competent. In a study, [86], Malle determined the moral acceptability of a set of propositions by requesting that participants judge their morality using the Moral Foundations Questionnaire [66], which measures the ethical principles of *harm*, *fairness* and *authority*. In this way, moral acceptability was determined by social agreement on the morality of propositions rather than by an external expert evaluation.

An exemplary implementation of a bottom-up approach is described in [96]. In that proposal, it is assumed that the agent will learn a model of societal preferences, and, when faced with a specific ethical dilemma at runtime,

efficiently aggregate those preferences to identify a desirable choice. This algorithm is based on a theory of voting rules. In the following, we reflect on the potential pitfalls of such an approach, which assumes that the choices of the crowd can be seen to reflect a system of ethics.

Bottom-up approaches to ethical reasoning aim to harness the wisdom of the crowd as a means to inform the ethical judgement of the agent. This view is in line with current approaches to AI, based on the development of models by observation of patterns in data. This approach assumes that a sufficiently large amount of data about ethical decisions and their results can be collected from a suitable set of subjects.

Reflection on the bottom-up approach

A fundamental premise of bottom-up approaches is the assumption that what is socially accepted – as evidenced in the data – is also what would be ethically acceptable. However, it is well known that sometimes stances that are *de facto* accepted are unacceptable by independent (moral and epistemic) standards and the available evidence. Conversely, there are other stances that are *de facto* not accepted but which are perfectly acceptable from a moral point of view. The difference between social acceptance and moral acceptability is that social acceptance is an empirical fact, whereas moral acceptability is an ethical judgement [123].

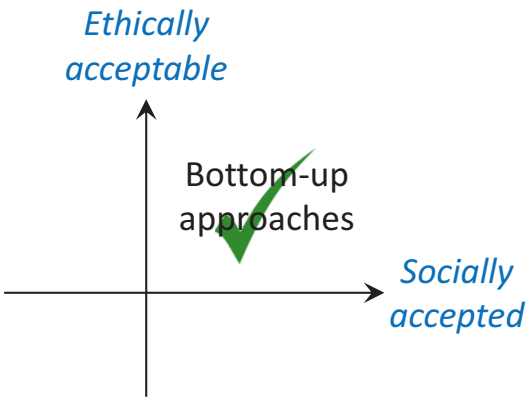

Figure 5.2: Bottom-up approaches assume alignment between social practice and ethics

In bottom-up approaches, the decision spectrum can be seen as a two-dimensional space along the *ethical acceptability* and the *social acceptance* axes (cf. Figure 5.2). Consensus on what is 'good' (acceptable and accepted) or 'bad' (unacceptable and unaccepted) behaviours or decisions is often easy

to achieve, but is always culture and context dependent. However, without extra incentives, the wisdom of the crowd can potentially lead to accepted but unacceptable decisions, i.e. those 'popular sins' such as tax avoidance or speeding, whereas morally acceptable stances are often not accepted by the crowd, e.g. due to perceived extra efforts or costs (examples of these are vegetarianism, the consumption of organic food, or accepting and supporting refugees and immigrants).

Moreover, each single opinion is often sustained by arguments of different acceptability, and even conflicting opinions can be sustained by equally acceptable ethical principles. For example, in discussions on whether to offer of fried foods and candy in school restaurants, both pro and con camps can base their opinions on equally 'good' moral arguments (healthy living and freedom of choice).

5.3.3 Hybrid Approaches

Hybrid approaches attempt to combine the benefits of top-down and bottom-up approaches as to ensure that ethical reasoning by AI systems is not only legally allowed but also socially accepted.

Gigerenzer [63] describes the nature of moral behaviour as the interplay between mind and environment. This view is based on pragmatic social heuristics rather than on moral rules or optimisation principles. In this view both nature and nurture are important in the shaping of moral behaviour. Extending this notion to ethical reasoning by AI systems, entailing a hybrid approach that combines programmed rules (nature) and context observations (nurture), is needed to implement ethical agents. Both nature and nurture need to be considered in conjunction, and not as an 'either/or' decision. This implies that neither a top-down approach, by means of programming, nor a bottom-up one, based on context, suffices to implement ethical decision making, but a combination of both approaches is needed.

One example of the application of a hybrid approach is the approach proposed by Conitzer and colleagues [30], which considers the integration of two (potentially complementary) paradigms for designing general moral decision-making methodologies: extending game-theoretic solution concepts to incorporate ethical aspects, and using machine learning on human-labelled instances to assemble an effective training set of labelled moral decision cases. Another example is the OracleAI system described in [8].

Reflection on the hybrid approach

By definition, hybrid approaches have the potential to exploit the positive aspects of the top-down and bottom-up approaches while avoiding their problems. As such, these may give a suitable way forward.

Recently, we have proposed a hybrid approach to ethical reasoning: MOOD [128]. MOOD is based on 'collective intelligence', that is, on bringing the knowledge and ideas of many minds together, but follows strict rules on how to elicit and aggregate these ideas. It also provides the means to make all design decisions explicit and queryable. MOOD aims to support free and constructive discussion and integration of several perspectives and, thereby, benefit several interest groups, not just the incumbent or 51% majority. In particular, it embeds the concepts of social acceptance and moral acceptability in the deliberation process.

Ethical acceptability concerns the fairness of decisions, but also the distributions of costs and benefits, the potential future harm to people and environment, risks and control mechanisms, and potential oppression and authority levels. The level of ethical acceptability score does not imply that an alternative should be selected or not, it merely provides insight into the ethical justness of the alternative. MOOD facilitates this type of debate taking into account the views of both the majority as well as the minority, and strives to be 'ethically just', that is, for stable and sustainable outcomes that are widely accepted.

5.4 Designing Artificial Moral Agents

Given the many complexities described in the previous sections, it should be clear that attempting to build systems that are able to take ethics into account in their reasoning is a tricky and complex endeavour. It is also one that should always start with the question: *"Should we develop such a system?"* Such systems, endowed with moral reasoning capabilities, are often referred to as *artificial moral agents*. This concept is currently popular both as a thought experiment and/or a real possibility [126], and has been reflected by many authors [132, 57, 6]. Even if fully ethical reasoning systems are not a realistic possibility, regardless of whether the agent is meant to be able to reason about the ethical aspects of its decisions, in many cases artificial systems are perceived by their users to take ethical decisions, or decisions that impact on one's ethics. It is therefore important to consider how to design AI systems such that their actions and decisions are aligned with societal and ethical principles. In this section, we provide guidelines to support the design of these systems towards an outcome that ensures responsible design. These guidelines should be seen as an extension to the Design for Values method sketched in Section 4.4.